US CLIVAR Climate Variability & Predictability

VARIATIONS

doi:10.5065/0DSY-WH17

Summer 2020 • Vol. 18, No. 2

New research on climate variability and change using initial-condition Large Ensembles

Guest Editors:
Clara Deser[1] and
Keith Rodgers[2,3]

[1]National Center for Atmospheric Research
[2]IBS Center for Climate Physics
[3]Pusan National University

A new thrust in climate and earth system modeling is to conduct an ensemble of simulations using the same model and radiative forcing protocol but varying the initial conditions. The resulting spread across the model ensemble, which is due solely to unpredictable internally-generated variability, places inherent limits on our ability to predict future climate change at regional and decadal scales. Such "initial-condition Large Ensembles" (LEs) also provide crucial context for understanding and interpreting the observational record, and foster robust model evaluation and inter-comparison by allowing the forced response to be separated from internally-generated variability. LEs also advance the study of extreme

Separating forced and unforced components of climate change: The utility of pattern recognition methods in Large Ensembles and observations

Robert C. J. Wills[1], Sebastian Sippel[2], and Elizabeth A. Barnes[3]

[1]University of Washington, USA
[2]ETH Zurich, Switzerland
[3]Colorado State University, USA

The identification of externally forced climate changes in the presence of internal climate variability is critical across many aspects of climate science, such as in attributing the causes of observed changes in weather and climate (Hegerl et al. 2007; Bindoff et al. 2013), understanding sources of uncertainty in future climate projections (Hawkins and Sutton 2009; Lehner et al. 2020), and improving predictions of internal climate variability on multi-year timescales (Meehl et al. 2014; Yeager et al. 2018). Single-model initial-condition large ensembles (SMILEs), ensembles of simulations with the same model and radiative forcing scenario but small differences in initial conditions, have become widely utilized for these purposes (Kay et al. 2015; Maher et al. 2019; Deser et al. 2020a). The forced response can be estimated

IN THIS ISSUE

events by providing a large number of samples of "rare" occurrences.

In this edition of Variations, we showcase new applications of LEs to the understanding of climate variability and change on regional and decadal scales. These articles, written by early-career researchers at the forefront of their fields, span a range of emerging topics including air quality and health, ocean biogeochemistry, best practices for evaluating models' internal variability, tools for quantifying forced changes in internal variability, and novel pattern recognition methods for detection and attribution. Many of these studies make use of a new multi-model archive of LEs conducted with CMIP5 models produced by the US CLIVAR Working Group on Large Ensembles, publicly available at http://www.cesm.ucar.edu/projects/community-projects/MMLEA/. This archive, along with LEs being conducted with CMIP6 models, herald a new era in climate science research and applications, and hold promise for new discoveries in the years to come.

US CLIVAR VARIATIONS

Editors:
Mike Patterson and Jennie Zhu
US CLIVAR Project Office
Washington, DC
usclivar.org

by averaging over the ensemble to remove internal variability that varies in phase between realizations. However, estimates of the forced response based on SMILEs are susceptible to any biases in the modeled forced response. It is therefore important to compare estimates of the forced response (e.g., a spatial pattern of change) across different models and to compare model-based estimates with estimates of the forced response from observations.

In order to estimate the forced climate response from observations, where only a single realization is available, a different approach is needed. Pattern recognition methods, including pattern-based statistical learning, and artificial intelligence, have particular utility because of the differences in spatial pattern between forced climate change and internal variability. For example, the climate response to anthropogenic greenhouse gas emissions is clearest at the global scale, where it manifests itself, for example, in warming of global-mean surface temperature. On the other hand, internal variability primarily redistributes heat between one region and another (or between the ocean and the atmosphere), such that it has a strong influence on regional climate but only a modest influence on global-mean surface temperature. Pattern differences can also extend to differences in vertical structure (e.g., Santer et al. 1996) or differences in the representation across multiple variables. Such differences in pattern between the signal and the noise have been exploited by a wide range of methods including standard detection and attribution methods (Hasselmann 1979; 1993; Bell 1986; Santer et al. 1995a; Hegerl et al. 1996; 2007; Bindoff et al. 2013), signal-to-noise-maximizing pattern analysis (Allen and Smith 1997; Schneider and Griffies 1999; Schneider and Held 2001; Ting et al. 2009; DelSole et al. 2011; Wills et al. 2018; 2020), dynamical adjustment (Wallace et al. 2012; Smoliak et al. 2015; Deser et al. 2016; Sippel et al. 2019), regularized regression (e.g., Sippel et al. 2020) and artificial neural networks (Barnes et al. 2019; 2020).

Large ensembles provide a testbed for methods to estimate the forced climate response from observations, but at the same time, the analysis of large ensembles can benefit from the use of pattern recognition methods. Within SMILES, statistical methods to estimate the forced response can be tested on individual ensemble members and then compared to the "true" forced response as estimated from the ensemble mean over a sufficiently large ensemble (e.g., Deser et al. 2016; Frankignoul et al. 2017; Sippel et al. 2019; Wills et al. 2020). However, for regional anomalies with a large amplitude of internal variability, the number of ensemble members needed to isolate the forced response with an ensemble average can become prohibitively large, with 50 or more ensemble members needed (Deser et al. 2012; Milinski et al. 2019). In much the same way that pattern recognition methods can be trained to estimate the forced response from observations, they can be trained to estimate the

forced response from a subset of ensemble members, reducing the number of ensemble members needed by up to a factor of ten in some cases (Wills et al. 2020).

Here, we give a brief example of the utility of pattern recognition methods and discuss how to best leverage a combination of pattern recognition methods and climate model ensembles to make progress on important questions related to separating forced and unforced components of climate change. Two research areas where the combination of pattern recognition methods and climate model ensembles have the potential to lead to major advances in understanding are in the analysis of structural uncertainty in climate projections and in characterizing the time evolving patterns of climate signal and climate noise.

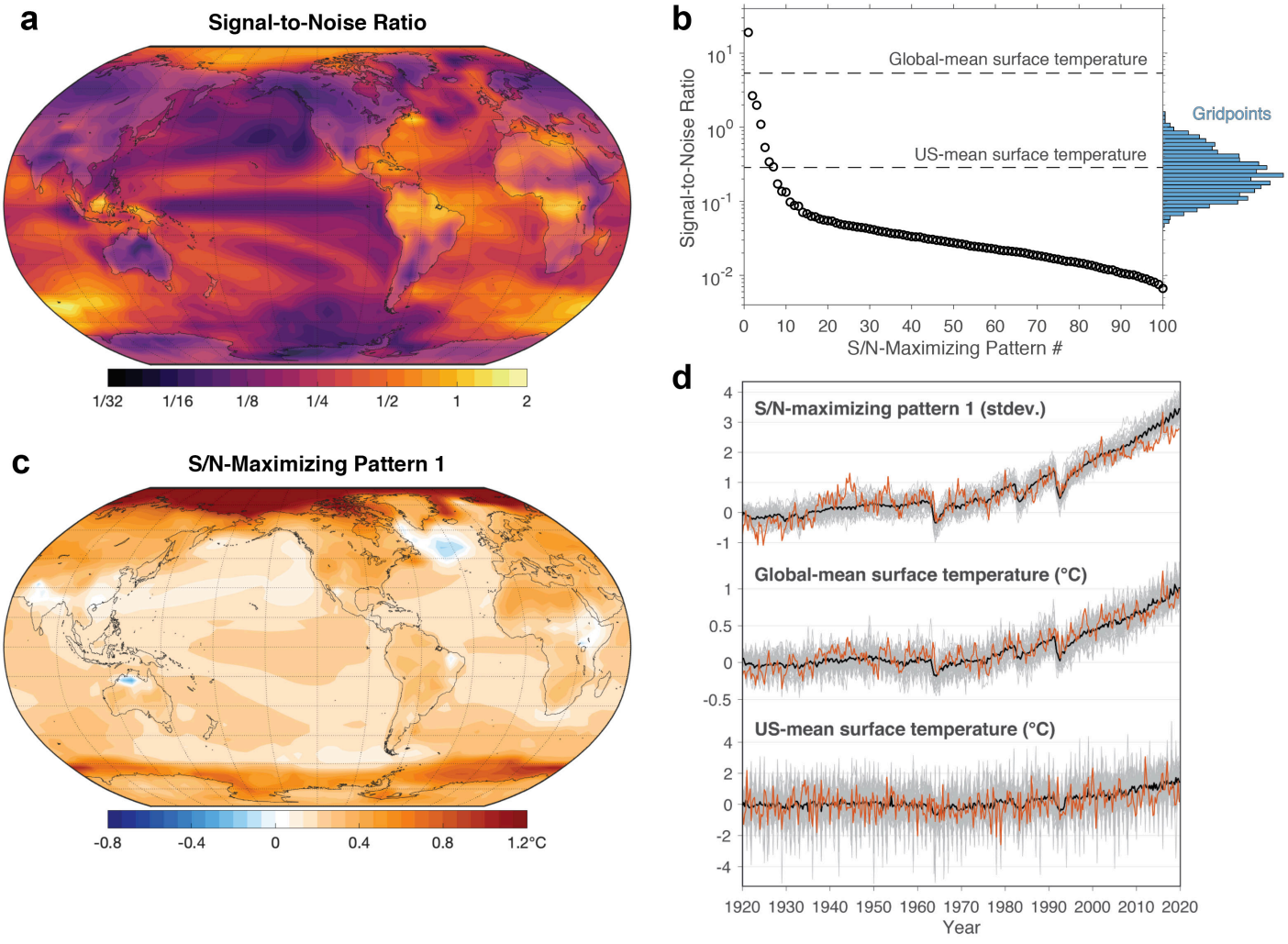


Figure 1. (a) Signal-to-noise ratio of seasonal (3-month-average) surface temperature anomalies in the CESM Large Ensemble (Kay et al. 2015) over the period 1920-2019. (b) Signal-to-noise ratio of signal-to-noise-maximizing (S/N-maximizing) patterns (Ting et al. 2009; Wills et al. 2020) of CESM-LE and comparison to range of signal-to-noise ratios for individual grid points (histogram on right axis), global-mean surface temperature, and US-mean surface temperature (dashed lines). (c) First S/N-maximizing pattern. (d) Time series of first S/N-maximizing pattern, global-mean surface temperature anomaly, and US-mean surface temperature anomaly. All anomalies are with respect to a 1920-1950 reference period. Grey lines show the time series in all 40 ensemble members, black lines show the ensemble mean, and orange lines show the HadCRUT4 observational temperature reconstruction (Cowtan and Way 2014). All analysis is done on a 5° longitude by 3.75° latitude grid.

The utility of pattern information

To illustrate the utility of pattern information, we compare the signal-to-noise ratio (SNR) of individual grid points within the CESM Large Ensemble (Kay et al. 2015) to that of the signal-to-noise-maximizing (S/N-maximizing) patterns, which are patterns that maximize the SNR within a truncated space of empirical orthogonal functions (Schneider and Griffies 1999; Ting et al. 2009; DelSole et al. 2011; Wills et al. 2020). Here, the signal is diagnosed using an ensemble average. The SNR is highest in the subtropical oceans and tropical land masses and lowest in the tropical Pacific and Northern Hemisphere midlatitudes (Figure 1a). However, even the grid points with the highest SNR (1.5 in the Indian Ocean sector of the Southern Ocean; 1.3 in Borneo) have SNRs over an order of magnitude lower than the leading S/N-maximizing pattern (SNR = 19, Figure 1b, c, d), which captures 71% of the ensemble-mean variance (i.e., the forced response). The S/N-maximizing pattern leverages information about the full spatial patterns of signal and noise, while the SNR computed for each grid point separately misses out on these spatial dependencies/relationships. The space of S/N-maximizing patterns is thus a useful basis for separating signal and noise within large ensembles and for reducing the dimensionality of the forced response. For example, Wills et al. (2020) truncate to just the leading ~10 S/N-maximizing patterns (which together capture 84% of the ensemble-mean variance) before taking an ensemble average, thereby reducing aliasing of internal variability onto the estimated forced response. A complimentary approach is that of dynamical adjustment (Wallace et al. 2012; Smoliak et al. 2015; Deser et al. 2016; Sippel et al. 2019), which identifies and removes patterns with low SNR based on their association with anomalies in the atmospheric circulation.

The utility of pattern information has long been recognized for the problem of detection and attribution of climate change (Hasselmann 1979; 1993; Bell 1986; Santer et al. 1995a; Hegerl et al. 1996), which seeks to detect a hypothesized climate response pattern (e.g., based on model simulations under specific forcing scenarios) within observations by identifying an "optimal fingerprint" that best distinguishes the response of interest from the background of internal variability. Unlike methods for characterizing the spatiotemporal evolution of the forced climate response (e.g., ensemble averaging, S/N-maximizing pattern analysis, dynamical adjustment), the goal of climate change detection is to determine whether a given forced response has occurred. The recent application of ideas from machine learning (e.g., ridge regression and neural networks), as well as the increasing magnitude of the climate signal itself, now allow detection of forced climate change from, for example, an individual year or day of data (Barnes et al. 2019; 2020; Sippel et al. 2020).

Fundamental to the climate change detection problem is the identification of a fingerprint or indicator pattern (e.g., Figure 2a) that represents how to weight observations in order to obtain a detection variable with high SNR. The same statistical machinery underlies S/N-maximizing pattern analysis, except that the response pattern (i.e., the S/N-maximizing pattern, Figure 1b) is determined empirically within a single climate-model dataset, rather than being imposed as a hypothesized response (i.e., no training is required). Each S/N-maximizing pattern has a corresponding fingerprint pattern, which is determined by multiplying the response pattern by the inverse covariance matrix (regularized by truncating in empirical orthogonal function space) and looks similar to those used in climate change detection applications (cf. Figure 2a; Hegerl et al. 1996; Sippel et al. 2020; Barnes et al. 2020). In our CESM Large Ensemble example, the projection of this fingerprint pattern onto observed temperatures (upper orange line in Figure 1d) shows a long-term trend in the S/N-maximizing pattern that emerges beyond the range of internal variability.

Analysis of structural uncertainty in climate projections

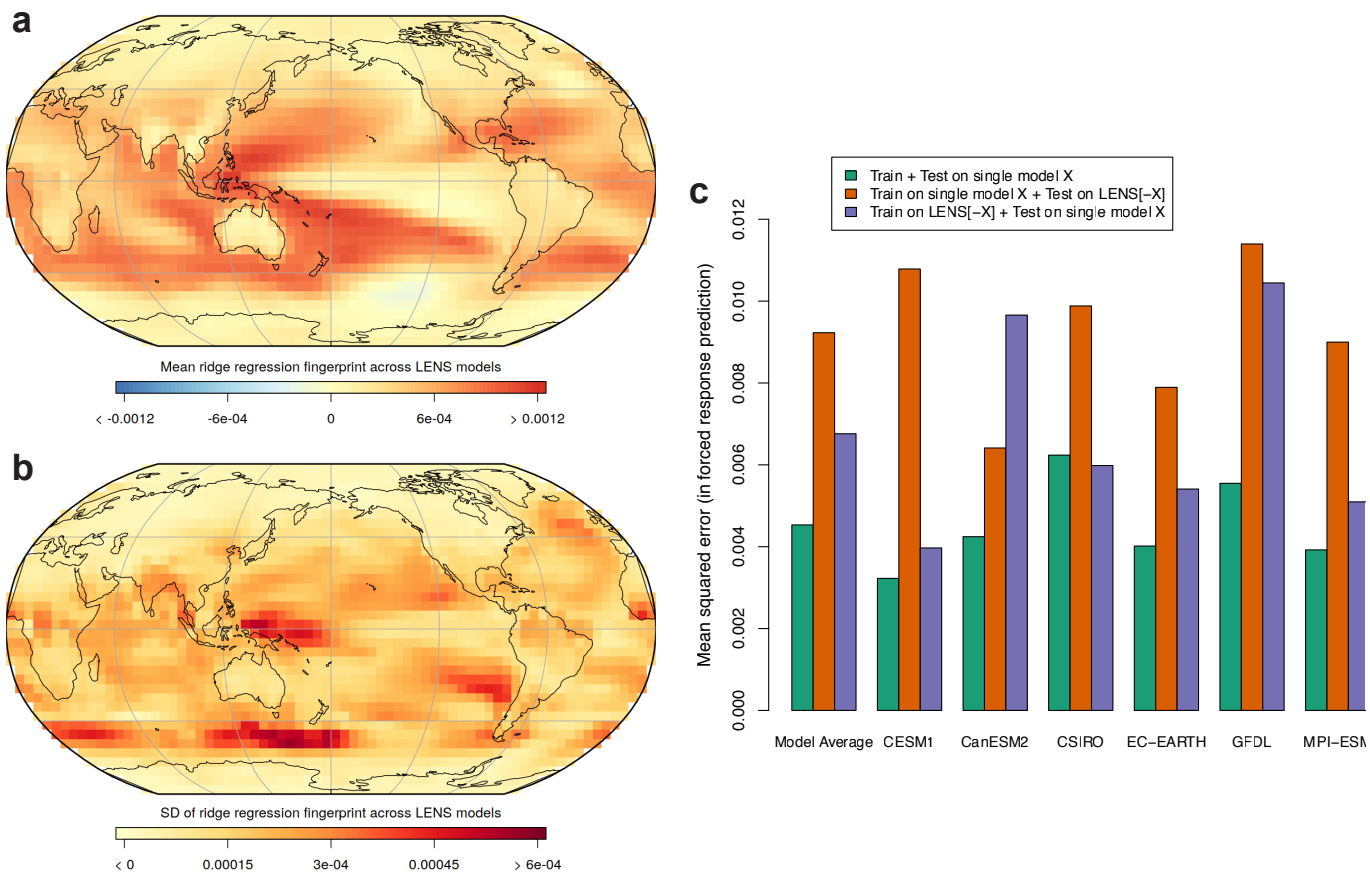


Figure 2. Illustration of the fingerprint for predicting the forced global-mean surface temperature response in the form of regression coefficients averaged over six different SMILEs for a fixed λ value (log10(λ) = 1.06). (b) Standard deviation of regression coefficients across different SMILEs, highlighting regions of model disagreement on the regression coefficients to optimally predict the forced global-mean surface temperature response. (c) Mean squared errors (in $(°C)^2$) calculated for the prediction of the annual-mean global-mean forced temperature response (from any year's spatial temperature pattern) for an average across models and from each individual SMILE. Colors indicate mean squared errors for (i) training and testing on a single SMILE (green bars), (ii) training on a single SMILE X and testing on all other models (LENS[-X], orange bars), and (iii) training on all but one model X (LENS[-X]) and testing on the SMILE X not used in training (blue bars). All analysis is done for the period 1951-2020 on a 5° longitude by 5° latitude grid. GFDL indicates a combination of the GFDL-CM3 and GFDL-ESM2M large ensembles.

Multi-model ensembles (e.g., the Coupled Model Intercomparison Project [CMIP]) have enabled substantial progress on the quantification of structural uncertainty in climate projections, i.e., differences in the forced climate response across models arising from differences in their formulation and tuning (e.g., Tebaldi and Knutti 2007). However, without multiple ensemble members from each of these models, structural uncertainty is partially confounded with uncertainty due to internal variability. Recently, a multi-model large ensemble archive (MMLEA) with seven different CMIP5-class models has been compiled (Deser et al. 2020a), enabling a clear separation of structural uncertainty and internal variability in these models. Pattern recognition methods provide utility for identifying differences in the forced response between these large ensembles, for comparing with additional models that have fewer ensemble members, and for formulating hypotheses about the spatiotemporal structure of the forced response (vs. internal variability) in order to detect the forced response in observations or in "holdout" climate models that have not been used in training.

One useful way to use pattern recognition methods in the analysis of multi-model ensembles is to apply the pattern recognition methods separately to each model (e.g., Wills et al. 2020), allowing identification of inter-model differences in the forced response, which can help identify and better understand model biases. Pattern recognition methods also reduce the number of ensemble members needed to isolate the forced response (Sippel et al. 2019; Wills et al. 2018; 2020), enabling analysis of structural uncertainty in climate projections in climate model ensembles ranging from the MMLEA to the broader CMIP archive.

For the purpose of detecting the forced climate response in observations, transferability of fingerprints extracted from imperfect models to observations becomes an important aspect. In detection and attribution, fingerprints of the expected forced response are extracted from models, and observations are only used in a second step (i.e., they are projected onto the fingerprints) to test whether the expected forced response can be detected in observations (e.g., Hegerl et al. 1996). The different patterns of forced response and internal variability across climate models (and imperfect knowledge of them in observations) offers an opportunity to utilize the heterogeneity provided by multi-model ensembles to increase the transferability of fingerprints to observations.

The benefit of pattern heterogeneity in multi-model ensembles for climate change detection is illustrated in Figure 2. We train a statistical model that predicts the forced (i.e., the ensemble mean) global-mean surface temperature response from the spatial pattern of surface temperatures. For the extraction of regression coefficients, we use ridge regression, a statistical learning technique able to deal with a large number of predictors, which avoids overfitting via a regularization parameter λ. For illustration, λ is fixed in this analysis to log10(λ) = 1.06, producing close to the minimum error when predicting a model not used in training (for details on ridge regression, see Hastie et al. 2009; for details on the climate application, see Sippel et al. 2020 and Barnes et al. 2020).

Training on each SMILE separately: Low error on training model, but poor transfer across models

We first train a ridge regression model for each of six SMILEs (CESM1, CanESM2, CSIRO, EC-Earth, GFDL, MPI-ESM) from the MMLEA. The average regression coefficients across the six individual models are illustrated in Figure 2a, indicating regions with high SNR to predict the forced response (cf. Figure 1a). These regions include tropical, subtropical, and some midlatitude regions, mainly in the world's oceans, but with the notable exception of the eastern equatorial Pacific, which exhibits El Niño variability that is uninformative for diagnosing forced climate change. Similar results were found, and discussed in more detail, in Barnes et al. (2019; 2020) and Sippel et al. (2020). Structural differences in the representation of the forced response and internal variability are evident in the variation in regression coefficients across models (Figure 2b). Regions of large disagreement include the western tropical Pacific, possibly indicative of Pacific cold tongue biases between models (Li et al. 2016), as well as regions in the Southern Ocean and the eastern subtropical Pacific.

We calculate the mean squared error (MSE) for the forced response prediction when the ridge regression model is trained and tested on the same model (green bars in Figure 2c). These errors are relatively small (~0.0045 $(°C)^2$, corresponding to a root mean squared error of ~0.067°C). However, because models differ in their representation of forced response and internal variability (Figure 2b), the fingerprint of SMILE X is expected to be suboptimal for predicting the forced response in models other than X (termed LENS[-X]). To test this, we show the MSE for predicting LENS[-X] using the fingerprint extracted only from SMILE X (orange bars in Figure 2c). As expected, the errors are higher when the fingerprint of a single model is used to predict the forced response in all other models (e.g., in the "all-model average", MSE is about double that of the single-model MSE; orange vs. green bars in Figure 2c).

Training on multiple SMILEs: improved transferability across models

Training on multiple models substantially improves the fingerprint transfer to an "unseen model." To demonstrate this, we train a second set of statistical models on all but one model (LENS[-X]) and evaluate the MSE for the model X not used in training. The prediction error is substantially reduced using the multi-model fingerprints instead of the single-model fingerprints (e.g., in the "all model average", MSE is reduced by around 30%; purple vs. orange bars in Figure 2c). Training across multiple models allows the algorithm to sample the heterogeneity of multi-model ensembles and improves the transferability of the resulting fingerprint to different, unseen models. If one adopts the assumption that the forced response and internal variability in observations may behave similarly to an "unseen model" in our example, then training across multiple models would be expected to improve the identification of the forced response in observations (e.g., Barnes et al. 2019, 2020; Sippel et al. 2020).

While an individual SMILE indeed provides a methodological testbed (as discussed in Deser et al. 2020a), evaluating expected error based on a single SMILE may not provide a representative evaluation of the transferability of a method or extracted feature/fingerprint to other models or to observations.

Time evolving patterns of climate signal and climate noise

The forced climate response is generally more complex than can be captured by a single spatial pattern that amplifies in time. This spatiotemporal complexity can arise, for example, due to the superposition of multiple types of radiative forcing (e.g., greenhouse gasses, anthropogenic aerosols, volcanic sulfur emissions, ozone) during the historical period. However, even the climate response to greenhouse gas forcing in isolation is thought to have a spatial pattern that evolves in time. In climate models, the changing pattern of warming in response to an abrupt increase in CO_2 concentrations plays an important role in the time evolution of global radiative feedbacks and global-mean surface temperature (Senior and Mitchell 2002; Armour et al. 2013; Andrews et al. 2015; Proistosescu and Huybers 2017; Dong et al. 2019).

In practice, the time evolving pattern of warming means that pattern recognition methods need to consider multiple forced response patterns. For example, in a S/N-maximizing pattern analysis of the CESM Large Ensemble over the period 1920-2019, the first ten patterns have SNRs that stand out from the continuum (Figure 1d). The higher order patterns (i.e., patterns 2-10) help to capture changes in seasonality and regional responses to forcing from anthropogenic aerosols and volcanic eruptions (Wills et al. 2020). While pattern recognition methods cannot by themselves distinguish between the effects of different types of radiative forcing within all-forcing simulations, they can help to characterize the differences between simulations with different radiative forcings, such as in large ensembles designed to isolate the influences of individual forcing agents (Santer et al. 2019; Deser et al. 2020b).

Formalisms for detecting multiple patterns of climate change and their evolution in time are not new (e.g., Santer et al. 1995a; Hegerl et al. 1997); these methods have been applied to increase confidence in the combined effect of greenhouse gasses and anthropogenic aerosols on observed temperature changes (see also Székely et al. 2019). Large ensembles provide new opportunities to apply these methods to quantify the impact of different forced response patterns over time, even in cases where a forced response is small relative to internal variability, and to quantify uncertainty in the time evolution of forced response patterns arising from internal variability (e.g., Santer et al. 2019).

Time evolution of the pattern of noise (variance) has received less attention than time evolution of the forced climate response. Partly this is due to modest changes in, for example, temperature variance over the historical period (Screen 2014), however, these changes will likely become larger in the future. Analogous to time evolving patterns of forced response, time evolving patterns of

noise can be addressed by including sufficient patterns to characterize the noise in both the reference climate and in the warmed climate (within existing methods such as optimal fingerprinting or S/N-maximizing patterns). However, nonlinear methods such as neural networks (Barnes et al. 2019; 2020) may be better suited to handle the coevolving patterns of climate signal and climate noise within climate change detection applications and could be explored in this context.

Discussion and conclusions

Forced climate change and internal variability have distinct spatial patterns. Pattern recognition methods can use this pattern information to separate forced and unforced components of climate change. Non-pattern-based methods for isolating the forced component of climate change, such as computing secular trends or regressing against global-mean surface temperature, do not take this spatial information into account, and thus, are less able to separate these components.

Large ensembles provide another tool to separate forced and unforced components of climate change, but analysis of large ensembles should still be designed in a way as to take advantage of pattern information. Recent work has suggested that pattern recognition methods can dramatically reduce the number of ensemble members needed to isolate the forced response (Wills et al. 2020), even allowing an approximate identification of the forced response within individual ensemble members or observations (Sippel et al. 2019; Wills et al. 2018; 2020).

The pattern recognition methods discussed in this article are by no means an exhaustive list. Diverse pattern-based methods should be explored in future work aiming to separate the forced and unforced components of climate change. In particular, the vast majority of pattern-based methods used in the literature assume a linear superposition of the various influences on climate. Future work should explore if and when nonlinear methods provide improved separation of the climate response from internal variability. Pattern recognition methods applied to large ensembles and observations have a strong potential to provide new frameworks for model evaluation and the analysis of structural uncertainty in climate projections, to improve the separation of forced and unforced components of climate change in observations, to separate the influences of different external forcings on climate changes, and to improve our understanding of the spatiotemporal evolution of climate change across multiple variables, from changes in average temperature to changes in climate extremes.

Acknowledgments

R. C. J. W. acknowledges funding provided by the National Science Foundation (Grant AGS-1929775). S. S. acknowledges funding provided by the Swiss Data Science Centre within the project 'Data Science-informed attribution of changes in the Hydrological cycle' (DASH, ID C17-01). E. A. B. is funded, in part, by the National Science Foundation (Grant AGS-1749261).

References

Allen, M. R., and L. A. Smith, 1997: Optimal filtering in singular spectrum analysis. *Phys. Lett. A*, **234**, 419–428, doi:10.1016/S0375-9601(97)00559-8.

Andrews, T., J. M. Gregory, and M. J. Webb, 2015: The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. *J. Climate*, **28**, 1630–1648, doi:10.1175/JCLI-D-14-00545.1.

Armour, K. C., C. M. Bitz, and G. H. Roe, 2013: Time-varying climate sensitivity from regional feedbacks. *J. Climate*, **26**, 4518–4534, doi:10.1175/JCLI-D-12-00544.1.

Barnes, E. A., J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2019: Viewing forced climate patterns through an AI lens. *Geophys. Res. Lett.*, **46**, 1389–13398, doi:10.1029/2019GL084944.

Barnes, E. A., B. Toms, J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, 2020: Indicator patterns of forced change learned by an artificial neural network. *J. Adv. Model. Earth Syst.*, doi:10.1029/2020MS002195.

Bell, T. L., 1986: Theory of optimal weighting of data to detect climatic change. *J. Atmos. Sci*., **43**, 1694–1710, doi:10.1175/1520-0469(1986)043<1694:TOOWOD>2.0.CO;2.

Bindoff N. L., and Coauthors, 2013: Detection and attribution of climate change: From global to regional. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Eds., Cambridge University Press, 867–952.

DelSole, T., M. K. Tippett, and J. Shukla, 2011: A significant component of unforced multidecadal variability in the recent acceleration of global warming. *J. Climate*, **24** , 909–926, doi:10.1175/2010JCLI3659.1.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn*., **38**, 527–546, doi:10.1007/s00382-010-0977-x.

Deser, C., L. Terray, and A. S. Phillips, 2016: Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications. *J. Climate*, **29**, 2237–2258, doi:10.1175/JCLI-D-15-0304.1.

Deser, C., and Coauthors, 2020a: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277-286, doi:10.1038/s41558-020-0731-2.

Deser, C., and Coauthors, 2020b: Isolating the evolving contributions of anthropogenic aerosols and greenhouse gases: A new CESM1 large ensemble community resource. *J. Climate*, In press, doi:10.1175/JCLI-D-20-0123.1.

Dong, Y., C. Proistosescu, K. C. Armour, and D. S. Battisti, 2019: Attributing historical and future evolution of radiative feedbacks to regional warming patterns using a Green's function approach: The preeminence of the western Pacific. *J. Climate*, **32**, 5471–5491, doi:10.1175/JCLI-D-18-0843.1.

Frankignoul, C., G. Gastineau, and Y.-O. Kwon, 2017: Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic multidecadal oscillation and the Pacific decadal oscillation. *J. Climate*, **30**, 9871–9895, doi:10.1175/JCLI-D-17-0009.1.

Hasselmann, K., 1979: *On the signal-to-noise problem in atmospheric response studies. Meteorology over the Tropical Oceans*, D. B. Shaw, Ed., Royal Meteorological Society, 251–259.

Hasselmann, K., 1993: Optimal fingerprints for the detection of time dependent climate change. *J. Climate*, **6**, 1957–1971, doi:10.1175/1520-0442(1993)006%3C1957:OFFTDO%3E2.0.CO;2.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 745 pp.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, **90**, 1095–1108, doi:10.1175/2009BAMS2607.1.

Hegerl G. C., J. von Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, 1996: Detecting greenhouse gas induced climate change with an optimal fingerprint method. *J. Climate*, **9**, 2281–2306, doi:10.1175/1520-0442(1996)009%3C2281:DGGICC%3E2.0.CO;2.

Hegerl, G. C., K. Hasselmann, U. Cubasch, J. F. B. Mitchell, E. Roeckner, R. Voss, and J. Waszkewitz, 1997: Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change. *Climate Dyn.*, **13**, 613-634, doi:10.1007/s003820050186.

Hegerl, G. C., and Coauthors, 2007: Understanding and attributing climate change. Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, S. Solomon et al., Eds., Cambridge University Press, 663-745.

Kay, J., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteorol. Soc*., **96**, 1333–1349, doi:10.1175/BAMS-D-13-00255.1.

Lehner, F., C. Deser, N. Maher, J. Marotzke, E. Fischer, L. Brunner, R. Knutti, and E. Hawkins, 2020: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth Syst. Dyn.*, **11**, 491-508, doi:10.5194/esd-11-491-2020.

Li, G., S.-P. Xie, Y. Du, and Y. Luo, 2016: Effects of excessive equatorial cold tongue bias on the projections of tropical Pacific climate change. Part I: The warming pattern in CMIP5 multi-model ensemble. *Climate Dyn.*, **47**, 3817–3831, doi:10.1007/s00382-016-3043-5.

Maher, N., and Coauthors, 2019: The Max Planck Institute grand ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst*., **11**, 2050-2069, doi:10.1029/2019MS001639.

Meehl, G. A., and Coauthors, 2014: Decadal climate prediction: An update from the trenches. *Bull. Amer. Meteorol. Soc.*, **95**, 243-267, doi:10.1175/BAMS-D-12-00241.1.

Milinski, S., N. Maher., and D. Olonscheck, 2019: How large does a large ensemble need to be?, *Earth Syst. Dyn. Discuss*., In review, doi:10.5194/esd-2019-70.

Proistosescu, C., and P. J. Huybers, 2017: Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Sci. Adv*., **3**, doi:10.1126/sciadv.1602821.

Santer, B. D., K. E. Taylor, T. M. Wigley, J. E. Penner, P. D. Jones, and U. Cubasch, 1995a: Towards the detection and attribution of an anthropogenic effect on climate. *Climate Dyn*., **12**, 77–100, doi:10.1007/BF00223722.

Santer B. D., U. Mikolajewicz, W. Brüggemann, U. Cubasch, K. Hasselmann, H. Höck, E. Maier-Reimer, and T. M. L. Wigley, 1995b: Ocean variability and its influence on the detectability of greenhouse warming signals. *J. Geophys. Res*., **100**, 10693–10725, doi:10.1029/95JC00683.

Santer, B. D. and Coauthors, 1996: A search for human influences on the thermal structure in the atmosphere. *Nature*, **382**, 39-46, doi:10.1038/382039a0.

Santer, B. D., J. C. Fyfe, S. Solomon, J. F. Painter, C. Bonfils, G. Pallotta, M. D. Zelinka, 2019: Quantifying stochastic uncertainty in detection time of human-caused climate signals. *Proc. Natl. Acad. Sci. U. S. A.*, **116**, 19821-19827, doi:10.1073/pnas.1904586116.

Schneider, T., and S. M. Griffies, 1999: A conceptual framework for predictability studies. *J. Climate*, **12**, 3133–3155, doi:10.1175/1520-0442(1999)012%3C3133:ACFFPS%3E2.0.CO;2.

Schneider, T., and I. M. Held, 2001: Discriminants of twentieth-century changes in earth surface temperatures. *J. Climate*, **14**, 249–254, doi:10.1175/1520-0442(2001)014<0249:LDOTCC>2.0.CO;2.

Screen J. A., 2014: Arctic amplification decreases temperature variance in northern mid- to high-latitudes. *Nat. Climate Change*, **4**, 577-582, doi:10.1038/nclimate2268.

Senior, C. A., and J. F. B. Mitchell, 2000: The time-dependence of climate sensitivity. *Geophys. Res. Lett.*, **27**, 2685–2688, doi:10.1029/2000GL011373.

Sippel, S., N. Meinshausen, A. Merrifield, F. Lehner, A. G. Pendergrass, E. Fischer, and R. Knutti, 2019: Uncovering the forced climate response from a single ensemble member using statistical learning. *J. Climate*, **32**, 5677–5699, doi:10.1175/JCLI-D-18-0882.1.

Sippel, S., N. Meinshausen, E. M. Fischer, E. Székely, and R. Knutti, 2020: Climate change now detectable from any single day of weather at global scale. *Nat. Climate Change*, **10**, 35–41, doi:10.1038/s41558-019-0666-7.

Smoliak, B. V., J. M. Wallace, P. Lin, and Q. Fu, 2015: Dynamical adjustment of the Northern Hemisphere surface air temperature field: Methodology and application to observations. *J. Climate*, **28**, 1613–1629, doi:10.1175/JCLI-D-14-00111.1.

Székely, E., S. Sippel, R. Knutti, G. Obozinski and N. Meinshausen, 2019: Proceedings of the 9th International Workshop on Climate Informatics: CI 2019. *9th International Workshop on Climate Informatics: CI 2019*; Boulder, CO, doi:10.5065/y82j-f154.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. R. Soc., A*, **365**, 2053-2075, doi:10.1098/rsta.2007.2076.

Ting, M., Y. Kushnir, R. Seager, and C. Li, 2009: Forced and internal twentieth-century SST trends in the North Atlantic. *J. Climate*, **22**, 1469–1481, doi:10.1175/2008JCLI2561.1.

Wallace, J. M., Q. Fu, B. V. Smoliak, P. Lin, and C. M. Johanson, 2012: Simulated versus observed patterns of warming over the extratropical northern hemisphere continents during the cold season. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 14337–14342, doi:10.1073/pnas.1204875109.

Wills, R. C., T. Schneider, J. M. Wallace, D. S. Battisti, and D. L. Hartmann, 2018: Disentangling global warming, multidecadal variability, and El Niño in Pacific temperatures. *Geophys. Res. Lett.*, **45**, 2487–2496, doi:10.1002/2017GL076327.

Wills, R. C. J., D. S. Battisti, K. C. Armour, T. Schneider, and C. Deser, 2020: Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations. *J. Climate*, In press, doi:10.1175/JCLI-D-19-0855.1.

Yeager, S. G., and Coauthors, 2018: Predicting near-term changes in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bull. Amer. Meteor. Soc.*, **99**, 1867–1886, doi:10.1175/BAMS-D-17-0098.1.

Internal climate variability and initial condition ensembles in air quality projections

James East and Fernando Garcia-Menendez

North Carolina State University

The use of initial-condition large ensembles (LEs) to examine internal variability in projections of climate change and its associated impacts has grown extensively in recent years. Although substantial research has investigated the impacts of climate change on air pollution under different warming scenarios (Jacob and Winner 2009; Fiore et al. 2015), LEs have not been used in these studies and are not yet available from fully coupled chemistry-climate models. However, there is evidence that internal variability can play a significant role in the projections, in particular when aiming to capture a "climate penalty" (i.e., a forced trend in the absence of anthropogenic precursor emission changes) on air quality and its associated health impacts. Here, we briefly review treatment of internal variability in prior efforts to simulate air quality under a changing climate. We then compare internal variability to other major sources of uncertainty in an ensemble simulation of climate change impacts on US air pollution, showing that internal variability can have an important effect on projections even at multidecadal and national scales. Finally, we discuss opportunities to develop more informative air quality projections by benefiting from LEs.

Internal variability in projections of air quality under climate change

In simulations of air quality under climate change, impacts are driven by pollutant emissions, the forced climate response, and unforced internal variability. Prior projections have focused on modeling the effects of changes in anthropogenic emissions, climate, or both (Jacob and Winner 2009; Fiore et al. 2015; Fu and Tian 2019). Most are based on a single model and single climate initialization. To account for internal variability, most studies use multiyear time slices when comparing present-day and future air quality. While the number of years varies across studies, time slices of ten or fewer years have been typically used to characterize air quality. These projections provide valuable insights into the potential impacts of climate variability and change on atmospheric composition and the mechanisms through which they can occur. However, the simulation length needed to detect a forced climate signal in ozone (O_3) and fine particulate matter ($PM_{2.5}$) concentrations above the noise of internal variability depends on the strength of this signal and magnitude of the noise (Barnes et al.

2016; Garcia-Menendez et al. 2017; Pienkosz et al. 2019). Although the analyses that use longer time slices can to an extent better capture internal variability, they still overlook low-frequency variability and the extent to which internal variability obscures the forced response depends on the study design. A projection's spatial scales, time horizon, pollutants of interest, and focus on the combined or separate effects of emissions and climate determine the feasibility of discerning a human-forced impact. Few studies have gone beyond the time slice approach to further investigate internal variability in air quality projections. Jimenez-Guerrero et al. (2011) apply empirical orthogonal functions to weigh internal variability in a transient 21st century simulation. Barnes et al. (2016) use multidecadal transient and control simulations to compare 20-year trends in surface ozone associated with internal variability to those driven by forced climate change or changes in anthropogenic emissions.

Perturbed initial condition ensembles are an important tool in climate modeling. Recently developed LEs show that without multiple initial conditions studies likely underestimate internal variability, which can lead to misattribution of impacts to the forced climate response (Deser et al. 2012). Only a small subset of air quality analyses, however, have included multiple climate initializations. Table 1 lists studies that use initial condition ensembles in projections of air quality under climate change. Projections based on a modeling framework that combines the MIT Integrated Global System Model and Community Atmosphere Model (MIT IGSM-CAM) include five initial conditions to account for internal variability in simulated impacts on air quality and several other sectors (Garcia-Menendez et al. 2015). Three-member initial condition ensembles are used in the future projections with the Geophysical Fluid Dynamics Laboratory chemistry-climate model, with 5 members used for the historical period (Turner et al., 2013; Clifton et al. 2014). Westervelt et al. (2016) further apply multiple linear regression to each of these ensemble members to breakdown projected impacts of climate change on fine particulate matter concentrations into sensitivities to different meteorological parameters. A Community Earth System Model (CESM1) initial condition ensemble of 15 members is used in projections of climate impacts on fine particulate matter pollution (Xu and Lamarque 2018). Some studies assess internal variability in meteorological conditions relevant to air quality without directly modeling atmospheric chemistry, and thus leverage some of the potential of initial condition ensembles while avoiding the computational costs of full chemistry simulations (e.g., Horton et al. 2012; Tai et al. 2012; Horton et al. 2014; Shen et al. 2017). Although here we focus on future climate, analyses examining present-day or past air quality have also sought to clarify the role of internal variability, with some similarly relying on initial condition ensembles (e.g., Turner et al., 2013; Lin et al. 2014; Hess et al. 2015; Callahan et al. 2019). Studies that include initial condition ensembles and identify specific mechanisms through which unforced climate variability affects air quality can provide further information about the model time scales needed to capture such variability (e.g., Lin et al. 2015; Callahan and Mankin 2020).

Multi-model ensembles with atmospheric chemistry include internal variability across members (e.g., Allen et al. 2016; Schnell et al. 2016; Silva et al. 2017) but this uncertainty is inherently mixed with model structural uncertainty. Differences across models complicate assessing internal variability specifically. While multi-model initiatives projecting climate and atmospheric composition have in some cases included as many as 15 models (e.g., Young et al. 2013), initial condition ensembles have not been explored within them (in part due to major constraints from limited computational resources). Still, discrepancies in projected pollutant concentrations across models are partially attributable to internal variability. Inconsistencies are especially apparent in projections of climate impacts on $PM_{2.5}$, including disagreements in the projected direction of climate-induced change (Von Schneidemesser et al. 2015).

Climate uncertainty in air quality projections

Three major sources contribute to uncertainty in climate projections: scenario uncertainty, model uncertainty,

Table 1. Studies projecting air quality under climate change that include transient (i.e., 21st Century) initial condition ensembles. Different greenhouse gas scenarios are considered.

Study	Initial condition members	Climate model	Atmospheric chemistry
Barnes et al. 2016	3	GFDL CM3	Coupled chemistry (gas-phase + aerosol)–climate
Callahan and Mankin 2020	35	CESM1 and 10 CMIP5 models	No interactive chemistry
Chen et al. 2018	3	GFDL CM3	Coupled chemistry (gas-phase + aerosol)–climate
Clifton, et al. 2014	3	GFDL CM3	Coupled chemistry (gas-phase + aerosol)–climate
Garcia-Menendez et al. 2015	5	MIT IGSM-CAM	Decoupled climate and chemistry
Garcia-Menendez et al., 2017	5	MIT IGSM-CAM	Decoupled climate and chemistry
Pienkosz et al., 2019	5	MIT IGSM-CAM	Decoupled climate and chemistry
Rieder et al. 2015	3	GFDL CM3	Coupled chemistry (gas-phase + aerosol)–climate
Rieder et al. 2018	3	GFDL CM3	Coupled chemistry (gas-phase + aerosol)–climate
Saari et al. 2019	5	MIT IGSM-CAM	Decoupled CAM-Chem chemistry
Turner et al. 2013	3	GFDL CM3	Coupled chemistry (gas-phase + aerosol)–climate
Tai et al. 2012	5	GISS GCM	No interactive chemistry
Westervelt et al. 2016	3	GFDL CM3	Coupled chemistry (gas-phase + aerosol)–climate
Xu and Lamarque 2018	15	CESM	Coupled aerosol–climate

and internal variability uncertainty (Hawkins and Sutton 2009). In projections of air quality under climate change, these uncertainties propagate to estimates of future pollution levels (which also depend on uncertain projections of air pollutant emissions). Here, we use an ensemble simulation of climate-induced impacts on 21st century US air quality (Garcia-Menendez et al. 2015) to illustrate how uncertainties associated with each of these sources compare. The simulations focus on climate-induced impacts on O_3 and $PM_{2.5}$ pollution, changes driven by variations in air pollution meteorology with emissions unchanged from start-of-

the century levels. Three scenarios of economic activity and greenhouse gas emissions are included. To capture model response uncertainty, simulations are conducted with climate sensitivities of 2.0, 3.0 and 4.5 °C. Air quality at the beginning, middle, and end of the 21st century is characterized by 30-year time slices. Five different climate initializations are used for each set of emissions scenario and climate sensitivity to further consider internal variability. The modeling framework based on the MIT IGSM-CAM and Community Atmosphere Model with atmospheric chemistry (CAM-Chem) is described in Garcia-Menendez et al. (2015). Further details about the methodology and projections are presented in Monier et al. (2015). Figure 1, reproduced from Pienkosz et al. (2019), shows the projected climate change impact on $PM_{2.5}$ at the end of the century under a no-policy Reference emissions scenario and a 3.0 °C climate sensitivity for each initial condition and individual years within the time slices. We weigh uncertainties as the mean range of 30-year-average O_3 and $PM_{2.5}$ surface concentrations (population-weighted over the US) for each source (emissions scenario, model response, and internal variability) across the ensemble, following the approach outlined by Monier et al. (2015) and building on Hawkins and Sutton (2009). Figure 2 shows uncertainties at the start, middle, and end of the 21st century in simulated daily-maximum 8-hour O_3 and $PM_{2.5}$. The uncertainties in modeled annual concentrations, rising as high as 3.8 ppb for O_3 and 1.2 μg m^{-3} for $PM_{2.5}$ by the end of the century, are similar in magnitude to ensemble-mean projected climate penalties on US air pollution (3.2 ppb and 1.5 μg m^{-3} under the Reference scenario and 3.0 °C climate sensitivity). Internal variability dominates uncertainty in projections prior to 2050. By midcentury, its influence on O_3 and $PM_{2.5}$ is comparable to that of emissions scenario and climate model response. While internal variability remains relatively constant throughout the simulations, the uncertainties associated with greenhouse gas

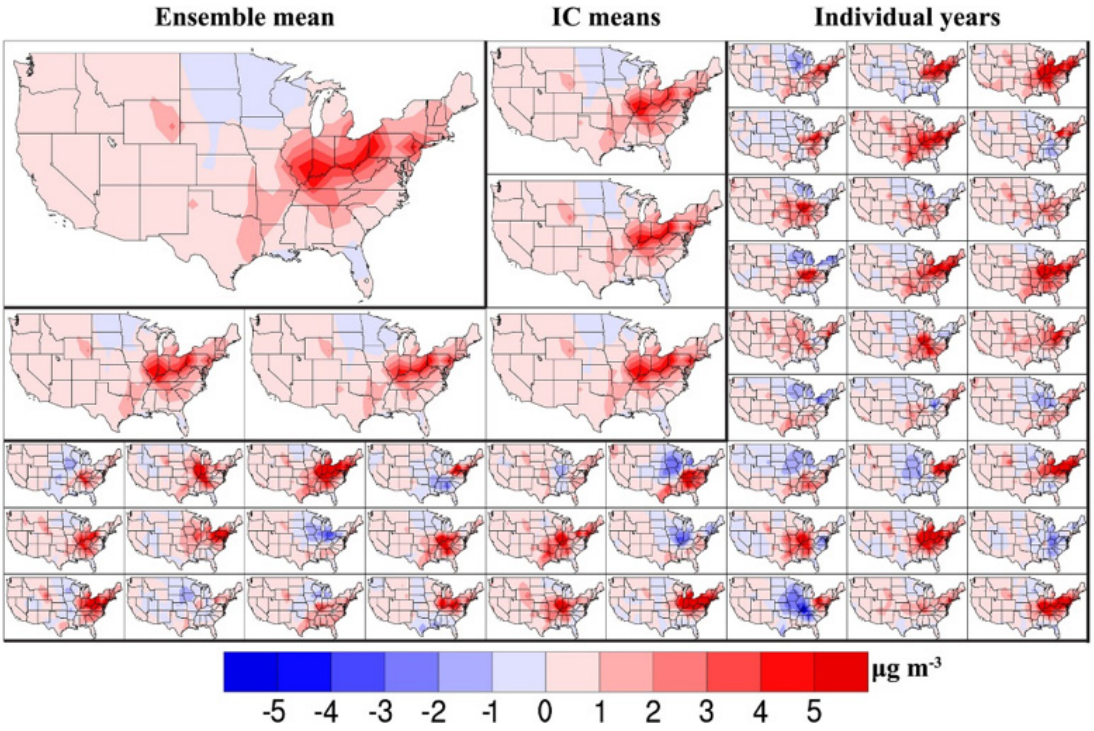


Figure 1. End of century ensemble-mean climate change impact on annual-average $PM_{2.5}$, relative to start of century, estimated under a no-policy Reference scenario and 3 °C climate sensitivity (top left panel). Mean (30-year) impacts estimated under each climate model initialization are shown in the five midsized panels. Samples of single year impacts, estimated as the difference between future-year and present-day concentrations, within the 30-year time slices are shown in the 45 smaller panels. (Pienkosz et al. 2019)

emissions and climate model response increase. Still, by the end of the century, uncertainty in projected air quality due to internal variability remains significant and equivalent to 20-30% of that imposed by emissions scenario, the prevailing uncertainty source. The contributions to uncertainty partially reflect those reported for mean surface air temperature and precipitation, with natural variability acting as the largest source early in the century and emissions scenario dominating by 2100 (Hawkins and Sutton 2009; Monier et al. 2015). The results also underscore the challenges imposed by internal variability on projecting a forced climate signal in air quality for time horizons extending over few decades. There are limitations to this assessment. Only three scenarios of greenhouse gas emissions and five initial conditions are included due to computational capacity. Variations in climate model response are represented applying a radiative adjustment method to a single model (Sokolov and Monier 2012). Despite these limitations, the analysis highlights the significance of internal variability in projections of air quality under climate change, even when examining multidecadal time slices at national scale.

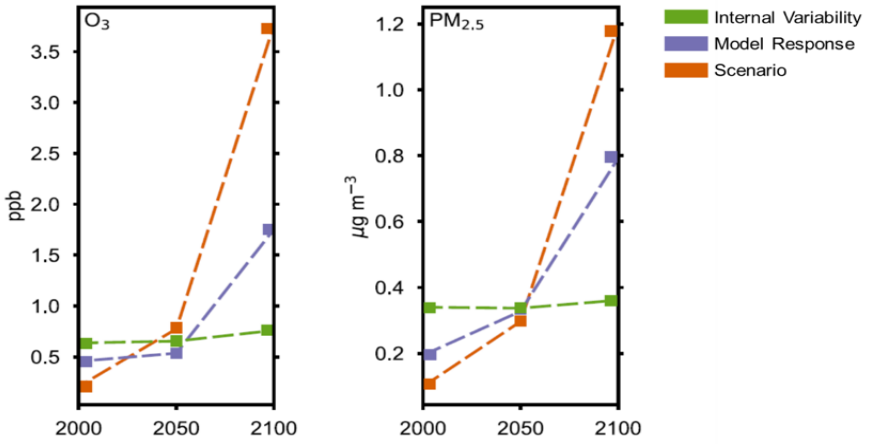


Figure 2. Mean (population weighted over the contiguous US) range of simulated annual-average surface 8-hour O_3 and $PM_{2.5}$ for each source of uncertainty at start, middle, and end of the 21st century.

Opportunities for LEs applications to air quality

LEs offer an exciting opportunity to explore the role of internal variability in air quality projections. Early efforts with small initial condition ensembles reveal that the approach can provide valuable insights (e.g., the studies in Table 1). Additionally, LEs have demonstrated the strong influence internal variability can have on meteorological variables recognized as important drivers of air pollution. However, the extent to which unforced internal variability impacts projections of atmospheric composition has not been carefully investigated. LEs provide a path to move beyond multiyear time slices, which by design are unable to capture internal climate variability on decadal and longer timescales, and better represent this uncertainty source. They may also help explain the differences among projections of climate-induced impacts on O_3, $PM_{2.5}$, and other harmful airborne pollutants.

There are several challenges associated with applying LEs to air quality. Among them is the added dimension of air pollutant and precursor emissions. Emission levels are anticipated to change significantly during the 21st century (Rao et al. 2017). While initial condition ensembles can help separate internal variability from human-caused changes in air quality, they cannot discern climate-induced impacts from those of varying anthropogenic pollutant emissions. Doing so requires simulations in which radiative forcing or air pollutant emissions are held static, increasing the amount of modeling necessary and introducing some new inconsistencies due to the dual role of some species as both climate forcers and important components of air pollution. Additionally, the dependence of forced changes and internal variability in atmospheric composition on air pollutant emissions levels has not been fully explored. LEs can add clarity to air quality projections aiming to simultaneously capture the impacts of climate change and anthropogenic emissions.

The high computational costs of full atmospheric chemistry simulations pose a challenge to developing LEs focused on air quality. When dynamical downscaling of climate fields is used to drive models with more complex chemical mechanisms at higher resolution, the requirements increase substantially. However, LEs are a reality in physical climate modeling and generating a single LE based on a fully coupled chemistry-climate model is viable with existing computational resources. Further, the latest global chemistry models now include representations of atmospheric interactions better able to simulate air pollution processes. For example, the CAM-Chem component of CESM2 includes an updated tropospheric chemistry mechanism (MOZART-T1), interactive aerosols, and 1° horizontal resolution (Emmons et al. 2020). Strategies to maximize forced signal detection (Brown-Steiner et al. 2018) can be used to explore internal variability within a single-model LE and selectively downscale a subset of the ensemble to drive regional climate or chemical transport models with more comprehensive treatments of air pollution chemistry and meteorology. The findings generated by a single-model LE can also motivate development of multi-model LEs, which would enable comparison of the internal variability generated across models, identification of biases in each model's representation of internal variability, and isolation of structural uncertainties (Deser et al. 2020). LEs with interactive atmospheric chemistry can be used to pursue new research directions, similar to those enabled by the development of ensembles focused on climate (Deser et al. 2020). The effects of internal variability on multiple aspects of air quality, including regulation attainment, extreme air pollution, natural emissions, and many others have not yet been investigated. Importantly, only a few studies have used small initial condition ensembles to project mortality and morbidity impacts of air pollution (Chen et al. 2018; Saari et al. 2019). However, propagation of this uncertainty to estimates of the future health burden of air pollution can have major implications. Initial research suggests that internal variability propagates beyond pollutant concentrations to the co-benefits of policies designed to slow climate change, and that uncertainty in these estimates can span a multi-trillion US$ range (Saari et al. 2019). Thus, correctly attributing impacts and characterizing uncertainty in air quality projections is an important research need. LEs offer a valuable tool to pursue this goal.

Acknowledgments

We greatly thank Arlene Fiore (Columbia University) and Daniel Horton (Northwestern University) for their valuable input and comments on this article.

References

Allen, R. J., W. Landuyt, and S. T. Rumbold, 2016: An increase in aerosol burden and radiative effects in a warmer world. *Nat. Climate Change*, **6,** 269-274, doi:10.1038/nclimate2827.

Barnes, E. A., A. M. Fiore, and L. W. Horowitz, 2016: Detection of trends in surface ozone in the presence of climate variability. *J. Geophys. Res,*, **121,** 6112–6129, doi:10.1002/2015JD024397.

Brown-Steiner, B., N. E. Selin, R. G. Prinn, E. Monier, S. Tilmes, L. Emmons, and F. Garcia-Menendez, 2018: Maximizing ozone signals among chemical, meteorological, and climatological variability. *Atmos. Chem. Phys.*, **18**, 8373-8388, doi:10.5194/acp-18-8373-2018.

Callahan, C. W., and J. S. Mankin, 2020: The influence of internal climate variability on projections of synoptically driven Beijing haze. *Geophys. Res. Lett.***, 47**, doi:10.1029/2020gl088548.

Callahan, C. W., J. L. Schnell, and D. E. Horton, 2019: Multi-index attribution of extreme winter air quality in Beijing, China. *J. Geophys.l Res.: Atmos.*, **124**, 4567–4583, doi:10.1029/2018JD029738.

Chen, K., A. M. Fiore, R. Chen, L. Jiang, B. Jones, A. Schneider, A. Peters, J. Bi, H. Kan, and P. L. Kinney, 2018: Future ozone-related acute excess mortality under climate and population change scenarios in China: A modeling study. *PloS Med***.,** 15, doi:10.1371/journal.pmed.1002598.

Clifton, O. E., A. M. Fiore, G. Correa, L. W Horowitz, and V. Naik, 2014: Twenty-first century reversal of the surface ozone seasonal cycle over the northeastern United States. *Geophys. Res. Lett.*, **41**, 7343-7350, doi:10.1002/2014GL061378.

Deser, C., R. Knutti, S. Solomon, and A. S. Phillips, 2012: Communication of the role of natural variability in future North American climate. *Nat. Climate Change*, **2**, 775-779, doi:10.1038/nclimate1562.

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277-286, doi:10.1038/s41558-020-0731-2.

Emmons, L. K., and Coauthors, 2020: The chemistry mechanism in the Community Earth System Model version 2 (CESM2). *J. Adv. Model. Earth Syst.*, **12**, doi:10.1029/2019MS001882.

Fiore, A. M., V. Naik, and E. M. Leibensperger, 2015: Air quality and climate connections. *J. Air Waste Manage. Assoc.*, **65**, 645-68, doi:10.1080/10962247.2015.1040526.

Fu, T. M., and H. Tian, 2019: Climate change penalty to ozone air quality: Review of current understandings and knowledge gaps. *Curr. Pollut. Rep*, *5*(3), 159-171, doi:10.1007/s40726-019-00115-6.

Garcia-Menendez, F., R. K. Saari, E. Monier, and N. E. Selin, 2015: U.S. air quality and health benefits from avoided climate change under greenhouse gas mitigation. *Environ. Sci. Technol.*, **49**, 7580–7588, doi:10.1021/acs.est.5b01324.

Garcia-Menendez, F., E. Monier, and N. E. Selin 2017: The role of natural variability in projections of climate change impacts on U.S. ozone pollution. *Geophys. Res. Lett.*, **44**, 2911–2921, doi:10.1002/2016GL071565.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, **90**, 1095–1107, doi:10.1175/2009BAMS2607.1.

Hess, P., D. Kinnison, and Q. Tang, 2015: Ensemble simulations of the role of the stratosphere in the attribution of northern extratropical tropospheric ozone variability. *Atmos. Chem. Phys.*, **15**, 2341–2365, doi:10.5194/acp-15-2341-2015.

Horton, D. E., Harshvardhan, and N. S. Diffenbaugh, 2012: Response of air stagnation frequency to anthropogenically enhanced radiative forcing. *Environ. Res Lett.*, **7**, doi:10.1088/1748-9326/7/4/044034.

Horton, D. E., C. B. Skinner, D. Singh, and N. S. Diffenbaugh, 2014: Occurrence and persistence of future atmospheric stagnation events. *Nat. climate change*, **4**, 698-703, doi:10.1038/nclimate2272.

Jacob, D. J., and D. A. Winner, 2009: Effect of climate change on air quality. *Atmos. Environ.*, **43**, 51–63, doi:10.1016/j.atmosenv.2008.09.051.

Jimenez-Guerrero, P., J. Jose Gomez-Navarro, S. Jerez, R. Lorente-Plazas, J. Andres Garcia-Valero, and J. Pedro Montavez, 2011: Isolating the effects of climate change in the variation of secondary inorganic aerosols (SIA) in Europe for the 21st century (1991-2100). *Atmos. Environ.,* **45**, 1059–1063, doi:10.1016/j.atmosenv.2010.11.022.

Lamarque, J. F., and Coauthors, 2013: The Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP): Overview and description of models, simulations and climate diagnostics. *Geosci. Model Dev.*, **6**, 179-206, doi:10.5194/gmd-6-179-2013.

Lin, M., L. W. Horowitz, S. J. Oltmans, A. M. Fiore, and S. Fan, 2014: Tropospheric ozone trends at Mauna Loa Observatory tied to decadal climate variability. *Nat. Geosci.*, **7**, 136–143, doi:10.1038/ngeo2066.

Lin, M., A. M. Fiore, L. W. Horowitz, A. O. Langford, S. J. Oltmans, D. Tarasick, and H. E. Rieder, 2015: Climate variability modulates western US ozone air quality in spring via deep stratospheric intrusions. *Nat. Commun.*, **6**, doi:10.1038/ncomms8105.

Monier, E., X. Gao, J. R. Scott, A. P. Sokolov, and C. A. Schlosser, 2015: A framework for modeling uncertainty in regional climate change. *Clim. Change*, **131**, 51–66, doi:10.1007/s10584-014-1112-5.

Pienkosz, B. D., R. K. Saari, E. Monier, and F. Garcia-Menendez, 2019: Natural variability in projections of climate change impacts on fine particulate matter pollution. *Earth's Future*, **7**, 762–770, doi:10.1029/2019EF001195.

Rao, S., and Coauthors, 2017: Future air pollution in the Shared Socio-economic Pathways. *Global Environ. Chang.,* **42**, 346-358, doi:10.1016/j.gloenvcha.2016.05.012

Rieder, H. E., A. M. Fiore, L. W. Horowitz, and V. Naik, 2015: 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.: Atmos.*, **120**, 784-800, doi:10.1002/2014JD022303.

Rieder, H. E., A. M. Fiore, O. E. Clifton, G. Correa, L. W. Horowitz, and V. Naik, 2018: Combining model projections with site-level observations to estimate changes in distributions and seasonality of ozone in surface air over the USA. *Atmos. Environ.*, **193**, 302-315, doi:10.1016/j.atmosenv.2018.07.042.

Saari, R. K., Y. Mei, E. Monier, and F. Garcia-Menendez, 2019: Effect of health-related uncertainty and natural variability on health impacts and cobenefits of climate policy. *Environ. Sci. Technol.*, **53**, 1098–11, doi:10.1021/acs.est.8b05094.

Schnell, J. L., M. J. Prather, B. Josse, V. Naik, L. W. Horowitz, G. Zeng, D. T. Shindell, and G. Faluvegi, 2016: Effect of climate change on surface ozone over North America, Europe, and East Asia. *Geophys. Res. Lett.*, **43**, 3509-3518, doi:10.1002/2016GL068060.

Shen, L., L. J. Mickley, and L. T. Murray, 2017: Influence of 2000–2050 climate change on particulate matter in the United States: Results from a new statistical model. *Atmos. Chem. Phys.*, **17**, doi:10.5194/acp-17-4355-20.

Silva, R. A., and Coauthors, 2017: Future global mortality from changes in air pollution attributable to climate change. *Nat. Climate Change*, **7**,647–652, doi:10.1038/nclimate3354.

Sokolov, A. P., and E. Monier, 2012: Changing the climate sensitivity of an atmospheric general circulation model through cloud radiative adjustment. *J. Climate*, **25**, 6567-6584, doi:10.1175/JCLI-D-11-00590.1.

Tai, A. P. K., L. J. Mickley, D. J. Jacob, E. M. Leibensperger, L. Zhang, J. A. Fisher, and H. O. T. Pye, 2012: Meteorological modes of variability for fine particulate matter (PM2.5) air quality in the United States: Implications for PM2.5 sensitivity to climate change. *Atmos. Chem. Phys.*, **12**, 3131–3145, doi:10.5194/acp-12-3131-2012.

Turner, A. J., A. M. Fiore, L. W. Horowitz, V. Naik, and M. Bauer, 2013: Summertime cyclones over the Great Lakes Storm Track from 1860-2100: Variability, trends, and association with ozone pollution. *Atmos. Chem. Phys.*, **12**, doi:10.5194/acp-13-565-2013.

Von Schneidemesser, E., and Coauthors: 2015: Chemistry and the linkages between air quality and climate change. *Chem. Rev.*, **11**, 3856-3897, doi:10.1021/acs.chemrev.5b00089.

Westervelt, D. M., L. W. Horowitz, V. Naik, A. P. K. Tai, A. M. Fiore, and D. L. Mauzerall, 2016: Quantifying PM2.5-meteorology sensitivities in a global climate model. *Atmos. Environ.*, **142**, 43–56, doi:10.1016/j.atmosenv.2016.07.040

Xu, Y., and J. F. Lamarque, 2018: Isolating the meteorologicalimpact of 21st century GHG warming on the removal and atmospheric loading of anthropogenic fine particulate matter pollution at global scale. *Earth's Future*, **6**, 428–440, doi:10.1002/2017EF000684.

Young, P. J., and Coauthors, 2013: Pre-industrial to end 21st century projections of tropospheric ozone from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP). *Atmos. Chem. Phys.*, **12**, doi:10.5194/acp-13-2063-2013.

Applications of Large Ensembles for marine biogeochemistry

Sarah Schlunegger[1], Yassir Eddebbar[2], Riley X. Brady[3], Friedrich A. Burger[4]

[1]Princeton University, USA
[2]University of California-San Diego, Scripps Institution of Oceanography, USA
[3]University of Colorado, USA
[4]University of Bern, Switzerland

Anthropogenic emissions and land-use change have driven atmospheric CO_2 concentrations, and consequently global temperatures, to rise. The ocean has removed 25-40% of anthropogenic carbon and more than 90% of anthropogenic heat from the atmosphere, thereby slowing the pace of global climate change (Frölicher et al. 2015; Gruber et al. 2019). This climate service provided by the ocean, however, does not come without penalty. Invading anthropogenic heat and CO_2 render the ocean warmer and more acidic, and reduce its oxygen content (Doney et al. 2009; Gruber 2011; Bopp et al. 2013). These impacts pose a threat to the ocean's capacity to continue sequestering anthropogenic CO_2 from the atmosphere and to remain a hospitable habitat for marine organisms and ecosystems.

Given the ocean's role as a large natural sink of anthropogenic carbon and the largest habitat on the planet, coordinated efforts are underway to observe our changing ocean and to predict its evolution over the coming centuries (e.g., Woods 1985; Takahashi et al. 2009; Wanninkhof et al. 2013; Friedlingstein et al. 2014). Early seminal work focused on understanding the global response of the ocean carbon cycle and climate to anthropogenic forcings (Oeschger et al. 1975; Siegenthaler 1983). Later, with increasing observational coverage, model complexity, and model resolution, the global response could be understood as the sum of heterogeneous regional responses (Maier-Reimer and Hasselmann 1987; Bacastow and Maier-Reimer 1987; Tans et al. 1990; Sarmiento and Orr 1992). The continuous expansion of our climate record has revealed substantial temporal variability in the ocean's physical and chemical state. This variability stems largely from internal variability in the climate system (Kosaka and Xie 2013; Landschützer et al. 2015), complicating the interpretation of the climate record and predictions of future change.

Assessing the anthropogenic influence on ocean biogeochemistry in multi-model studies (e.g., through the Coupled Model Intercomparison Project [CMIP])

has been challenging due to insufficient model member ensembles, which confounds uncertainty due to internal variability with that of model structure (Hawkins and Sutton 2009). This is especially relevant for key biogeochemical variables such as net primary productivity (NPP) and dissolved oxygen (O_2), which showcase large inter-model differences and biases as well as substantial internal variability (Bopp et al. 2013; Cabré et al. 2015; Ito et al. 2017; Long et al. 2016).

Initial Condition Large Ensemble experiments (LEs) conducted with Earth system models (ESMs) represent an important advancement in attributing historical trends and projecting future natural and anthropogenic changes in ocean biogeochemistry. In this article, we survey recent applications of LEs to ocean biogeochemistry for addressing questions about timescales of anthropogenic change, mechanisms of variability, and ecological vulnerabilities.

Survey of BGC Applications

Emergence

Time of emergence (ToE) represents the statistical point in time at which an anthropogenic trend (the signal) statistically exceeds internal climate variability (the noise). ToE is a multi-purpose metric which informs (1) the timescales over which impacts of climate change might occur, as this is when the anthropogenic trend exceeds the natural variability to which organisms and systems are adapted, and (2) the implementation and interpretation of climate observations, as statistical emergence is a necessary prerequisite for detection of anthropogenic trends. LE experiments with ESMs, in addition to multi-model ensembles (e.g., CMIP5), have been used to show that biogeochemical properties and processes in the ocean have disparate times of emergence that range from under a decade—for properties like surface ocean pH—to over a century—for processes like NPP (Figure 1; Rodgers et al. 2015; McKinley et al. 2016; Frölicher et al. 2016; Lovenduski et al. 2016; Long et al. 2016; Henson et al. 2017; Carter et al. 2016; Li et al. 2017; Schlunegger et al. 2019; Schlunegger et al., 2020).

LEs, in tandem with sensitivity studies, reveal that this chronology of emergence is a consequence of the differing timescales associated with the rapid chemical invasion of anthropogenic CO_2 into the ocean and long-term warming effects on ocean biogeochemistry. The impacts on carbonate chemistry and calcification emerge first, followed by the slow adjustment of ocean ventilation to warming and freshening, and lastly, by changes to nutrient supply and biological activity (Schlunegger et al. 2019). Emergence timescales are not only disparate across different biogeochemical variables, but can also differ substantially by region for the same variable (e.g., Figure 1b-d). For example, emergence of thermocline O_2 in the Community Earth System Model Large Ensemble (CESM-LE; Kay et al. 2015) can vary from a few decades in certain regions to more than a century in others, reflecting distinct regional sensitivities to governing modes of variability and anthropogenic forcing (Long et al. 2016).

The complex spatial composition and long emergence timescales of many important biogeochemical properties underscores the need for sustained, large-scale observations of the ocean in order to monitor and adapt to anthropogenic change.

Mechanisms of variability

The delayed emergence of anthropogenic trends in certain variables (e.g., O_2 and NPP) in comparison to early emergence in others (e.g., pH and pCO_2) also reflects key differences in how natural climate variability modulates biogeochemical cycles. This variability can be driven by a range of processes, such as internal modes of climate variability (e.g., El Niño–Southern Oscillation (ENSO), North Pacific Gyre Oscillation (NPGO), Southern Annular Mode (SAM)) as well as external drivers, such as volcanic eruptions. In contrast to multi-model ensembles which obfuscate uncertainty

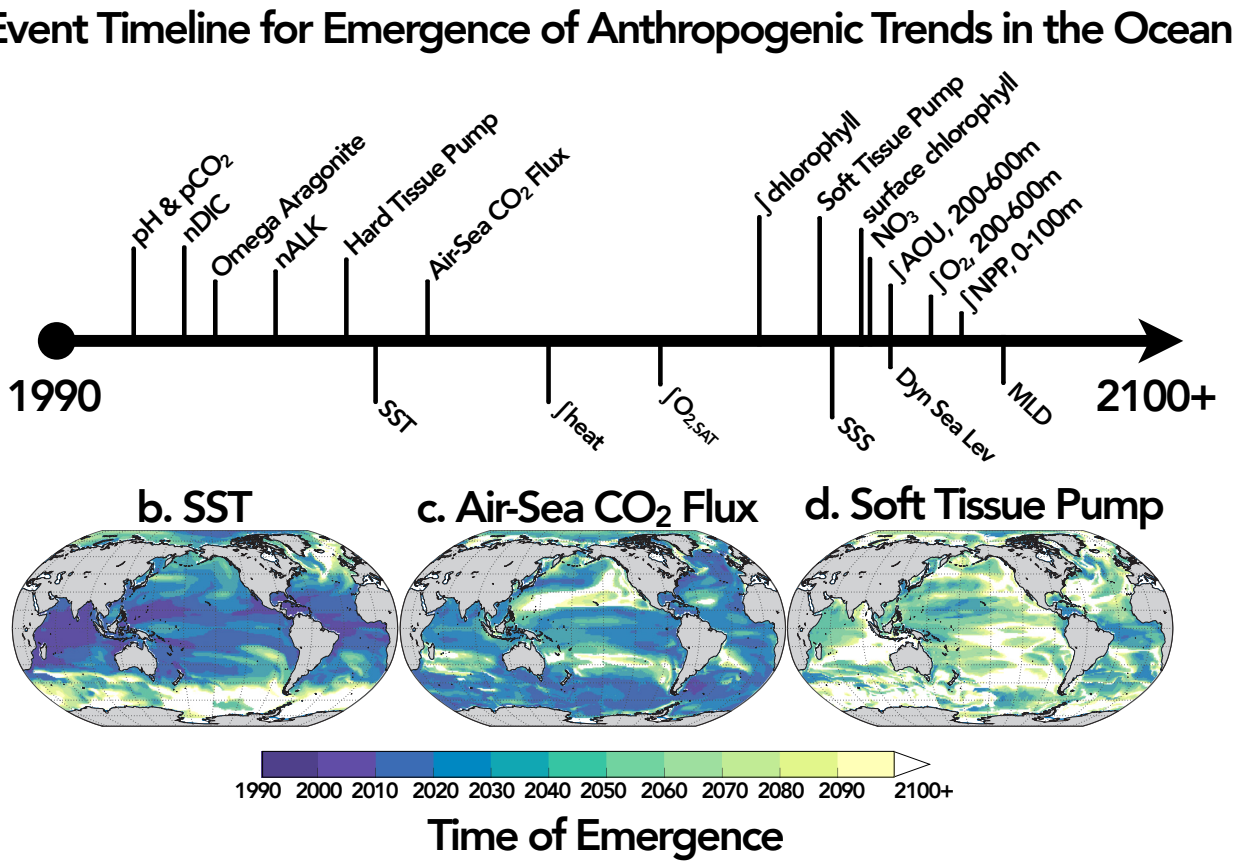


Figure 1. Time of Emergence (ToE) for anthropogenic trends in marine biogeochemistry for the GFDL-ESM2M RCP8.5 Large Ensemble Experiment. ToE calculations are referenced to year 1990, as this is the approximate beginning of the ocean-observing era and therefore the start of the reference period from which contemporary anthropogenic trends can emerge. The timeline in panel (a) indicates the year at which 50% of the global ocean area has an emergent anthropogenic trend for the given biogeochemical variable. The progression of emergence is a consequence of the initial, rapid invasion of anthropogenic CO_2 and its impact on carbonate chemistry and calcification, followed by the slower response of the climate system and ocean to anthropogenic radiative forcing, and subsequent changes in ventilation, nutrient supply, and biological activity. The maps show the spatial patterns of emergence for (b) sea surface temperatures (SST), (c) air-sea CO2 fluxes and (d) the soft tissue pump, defined as the export of particulate organic matter at 100 meters depth. Other variable definitions include salinity normalized Dissolved Inorganic Carbon (nDIC) and Alkalinity (nALK), the equilibrium saturation concentration of O_2 ($O_{2,SAT}$) integrated from 200-600 m, Apparent Oxygen Utilization (AOU), Sea Surface Salinity (SSS), Dynamic Sea Level (Dyn Sea Lev) and Mixed Layer Depth (MLD).

due to internal variability and model structure, the LE framework allows one to isolate the forced response of ocean biogeochemistry by averaging over many realizations of the same climate model with the same external forcing (Deser et al. 2012; Kay et al 2015). This response includes both the forced trajectory of the system (ensemble mean in bold lines in Figure 2a-c), as well as its internal variability (ensemble spread outlined by the individual realizations in thin lines Figure 2a-c).

Internal variability

The response of ocean biogeochemistry to internal variability exhibits large spatial heterogeneity (Figure 2d-f). Some regions emerge as hot spots for internal variability across all biogeochemical variables, such as the equatorial Pacific, Eastern Boundary Upwelling Systems (EBUS), Southern Ocean, and subpolar North Atlantic and North Pacific Oceans. This is due to the sensitivity of these regions to internal modes of climate

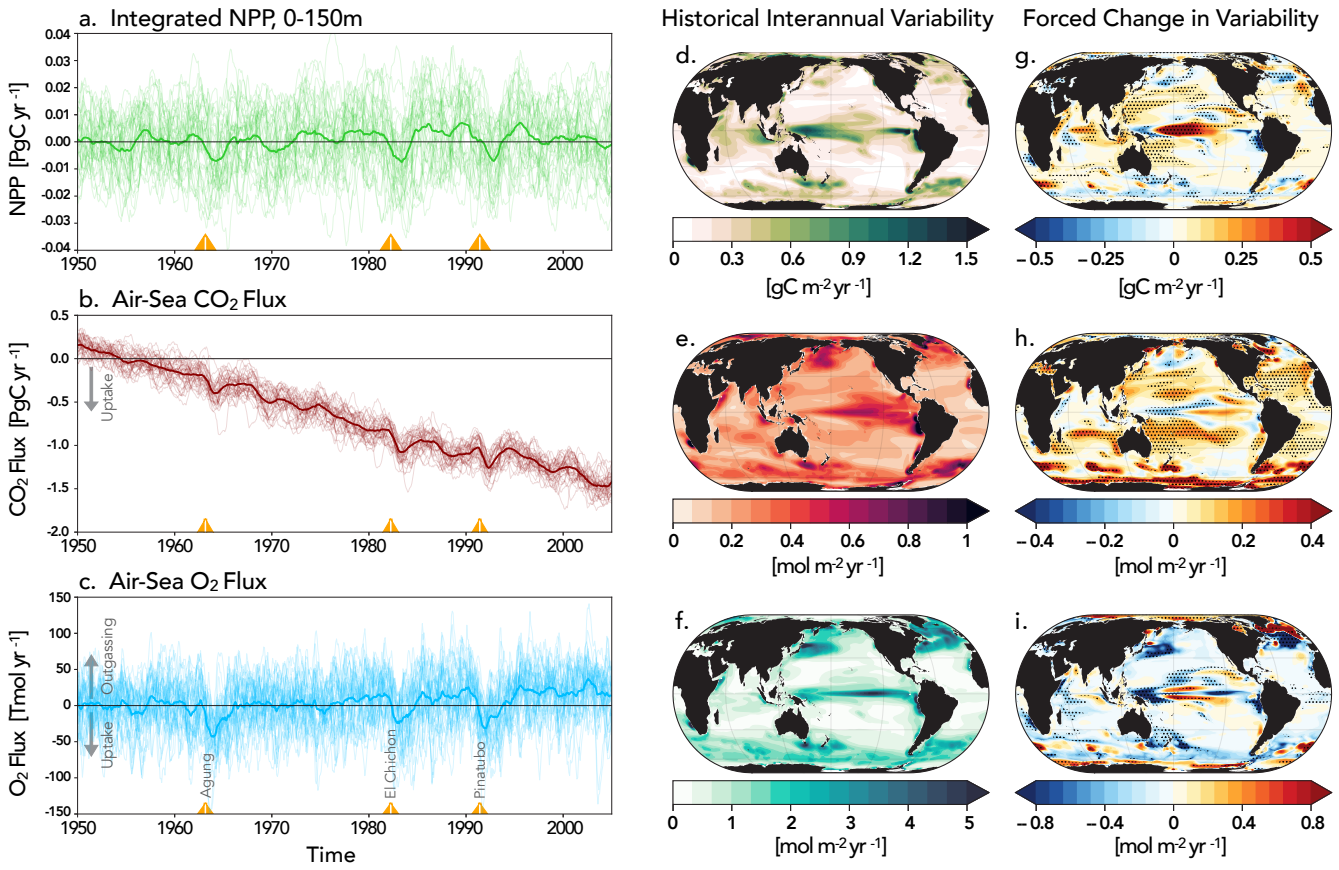


Figure 2. Evolution of key biogeochemical variables over the historical period, including a) net primary production integrated over the upper 150 m, b) air-sea flux of CO_2, and c) air-sea flux of O_2 (negative values denote uptake). The solid lines show the ensemble mean (forced response), while the thin lines show the individual 34 ensemble members of CESM-LE (internal variability), all smoothed with a 12-month running mean and adjusted by the 1950 decadal mean as a reference period. The three large historical volcanic eruptions (Mount Agung 1963, El Chichón 1982, and Mount Pinatubo 1991) are marked by triangles in panel c. (d-f) Historical unforced interannual variability of biogeochemical variables, computed by taking the ensemble mean standard deviation over 1971-2000 after removing the ensemble mean forced response. (g-i) Forced change in interannual variability between 1971-2000 and 2071-2100, represented by the ensemble mean of the individual changes in interannual variability for each ensemble member. Stippling denotes statistical significance at the 95% confidence level, where the signal-to-noise ratio is greater than two. Here, the signal is the ensemble mean change, while the noise is the ensemble spread of the change. Similar changes to variability are found by examining the evolution of ensemble spread over time.

variability, such as ENSO in the tropics (e.g., McKinley et al. 2003; Long et al. 2013; Eddebbar et al. 2017; McKinley et al. 2017), and other modes of variability at higher latitudes (e.g., SAM, Lovenduski et al. 2007; Verdy et al. 2007; NPGO, Di Lorenzo et al. 2008; Brady et al. 2019). Differences in the spatial structure of internal variability of these biogeochemical variables are also evident, such as the intense response of the western equatorial Pacific for NPP and the western boundary currents and equatorial Pacific for air-sea O_2 flux.

An advantage of using the LE framework is its repeated sampling of the historical climate with evolving external forcing. This allows one to assess the robustness of a relationship between modes of climate variability and ocean biogeochemistry over a given time period. Brady et al. (2019) leveraged this to identify the major modes of climate variability that are associated with anomalous air-sea CO_2 flux in EBUS. They performed decompositions of the response of driver variables of air-sea CO_2 flux to the given mode of variability for each ensemble member.

They found that anomalous air-sea CO_2 flux in EBUS is largely driven by the opposing response of CO_2 solubility and dissolved inorganic carbon (DIC) to modes of climate variability. For example, during a positive NPGO event in the California Current, upwelling is intensified by enhanced gyre circulation, which cools SSTs and thus increases the ingassing pressure, while enhanced transport of DIC from the Alaskan Gyre intensifies the outgassing pressure (Brady et al. 2019). The large number of realizations from the LE places robust uncertainty bounds on the response of the given variable to modes of climate variability over the historical period.

While internal variability can be isolated through a preindustrial control run with constant CO_2 forcing (e.g., Frölicher et al. 2009; Long et al. 2013), this method does not address changes to internal variability due to external forcing. One requires many realizations of the same climate model to isolate the evolving response of internal variability to anthropogenic forcing. Figure 2g-i highlights changes to internal variability in NPP, air-sea O_2 flux, and air-sea CO_2 flux on the interannual scale under the RCP8.5 scenario in CESM-LE. An increase in internal variability due to external forcing can also be seen for seawater acidity [H+] (Figure 3b).

Changes to internal variability in biogeochemistry can be driven by both (1) changes in the variability of the physical climate system and (2) by the non-linear sensitivities of chemical and biological processes to the mean physical and chemical state of the ocean. Physically driven changes in biogeochemical variability are expected to arise from modifications to the frequency and expression of climate modes like ENSO (Timmermann et al. 1999; Stevenson et al. 2019), or from other processes such as the poleward expansion of Hadley Cells (Yin et al. 2005), sea ice melt (Bates et al. 2006) and warming-induced stratification (Freeman et al. 2018). Even without modifications to the variability of the physical climate, rising ocean temperatures and increasing concentrations of DIC result in a more sensitive, and thus more variable air-sea CO_2 flux (e.g. Fassbender et al. 2018; Landschützer et al. 2018).

Forced variability

Volcanic eruptions induce additional variability in the climate system, further challenging the attribution of anthropogenic trends in ocean biogeochemistry. These eruptions can influence ocean biogeochemical dynamics

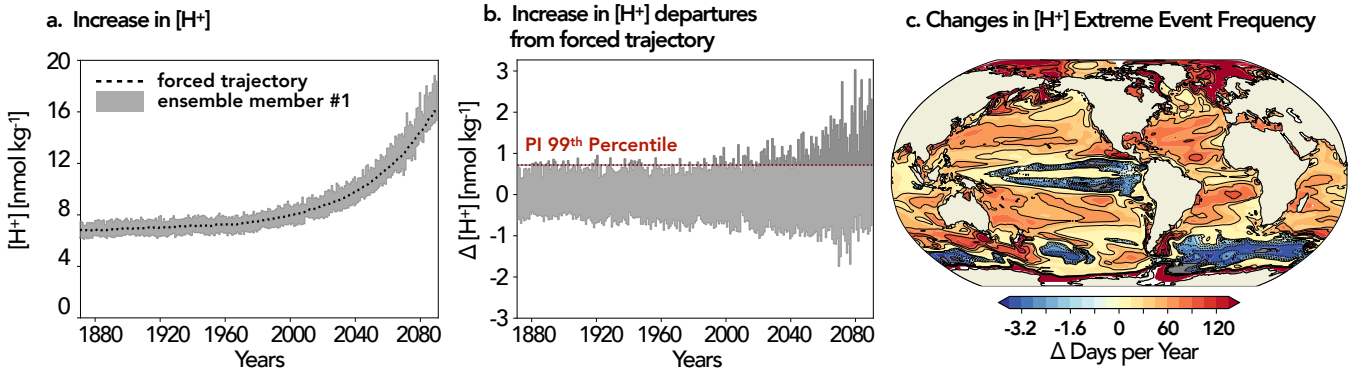


Figure 3. (a) Yearly range of daily-mean surface seawater acidity [H+] for a grid cell at 40°N and 30°W over the historical and RCP8.5 period for one member of an ensemble simulated with the GFDL ESM2M model. The black dotted line indicates the ensemble mean that represents the forced trajectory in [H+]. (b) The yearly range of departures from the forced trajectory for the data shown in a). The red dotted line is the extreme event threshold, defined as the 99th percentile of departures in a preindustrial (PI) simulation. (c) The change in surface [H+] extreme event frequency between the preindustrial and the ensemble average over the years 2081-2100 under RCP8.5.

through their direct aerosol radiative effects on sea surface temperatures (SSTs) which control gas solubility and air-sea gas exchange (Frölicher et al. 2009; 2011), as well as their indirect radiative effects on ocean circulation by modulating spatial gradients of SSTs and ocean-atmosphere interactions (Zanchettin et al. 2012). Isolating the volcanic signature on ocean biogeochemistry has long been challenging in the presence of obfuscating internal variability and the small number of ensemble members in previous model studies (Frölicher et al. 2009; 2011).

The LE framework, which simulates the volcanic aerosol effects on climate, is ideal for examining volcanic effects on ocean biogeochemical dynamics across models and volcanic forcing datasets for exploring governing mechanisms. Figure 2a-c illustrates the biogeochemical impacts of the Mount Agung (1963), El Chichón (1982), and Mount Pinatubo (1991) eruptions in the CESM-LE. These tropical eruptions drive reduced NPP and enhanced oceanic uptake of O_2 and CO_2 that penetrate into the ocean's interior, where they persist for several years (Eddebbar et al. 2019). The spatial patterns and mechanisms driving this oceanic uptake of oxygen and carbon differ considerably in the LE. Whereas anomalous CO_2 uptake occurs mainly at lower latitudes, driven by reduced upwelling of DIC-rich waters and surface cooling, O_2 drawdown occurs primarily at higher latitudes, where colder winters following eruptions induce vigorous uptake of O_2 (Eddebbar et al. 2019).

The LE experiment suggests that tropical eruptions act to temporarily slow down the progression of ocean deoxygenation due to ocean warming (Keeling et al. 2010) and to briefly intensify the oceanic uptake of anthropogenic carbon and induce interannual to decadal variability in the global ocean carbon sink (McKinley et al. 2020). This complex sensitivity of ocean biogeochemistry to radiative energy perturbations should be taken into account in attribution studies of the observed variability in the oceanic oxygen and carbon cycles (Ito et al. 2017; Landschützer et al. 2015; McKinley et al. 2017; 2020).

Ecological applications

LE experiments can be used to study the ecological impacts of changes in physical and biogeochemical ocean variables. This includes the use of LE experiments to detect changes in extreme events (Burger et al., 2020) and the crossing of biological thresholds (Negrete-Garcia et al. 2019), as well as to force ecological models to assess changes therein (Krumhardt et al. 2017; Cheung and Frölicher 2020). LEs are a particularly valuable tool for analyzing extreme events, since they provide enough data to study rare events and to track changes in extreme event statistics over transient simulations. Furthermore, LEs are useful for understanding variance changes in non-linear systems like carbonate chemistry, as the LE mean allows for the forced, non-linear trend in the mean-state to be removed without assumptions or statistical fitting methods. Here, we review a selection of studies that utilize LEs to assess the ecological impacts of anthropogenic forcing.

Burger et al. (2020) analyze how the occurrence and the characteristics of extreme events in [H+] change with anthropogenic CO_2 emissions within an ensemble simulation. Extreme events are defined with respect to a moving baseline, which is identified by the changing ensemble mean (Figure 3a). The study demonstrates that, in addition to the large increases in mean acidity, increasing variability leads to larger departures from the evolving mean states (Figure 3b) and to increased occurrence of extremes (Figure 3c). Cheung and Frölicher (2020) detect marine heatwaves in a large ensemble simulation and study their impacts on fish stocks in the Northeast Pacific using a fish impact model. They show that the impact of marine heatwaves on biomass decreases for some species could be several times higher than the decrease per decade due to global warming.

Negrete-Garcia et al. (2019) detect sudden shoaling events of the aragonite saturation horizon in the Southern Ocean within the CESM-LE. The aragonite

saturation horizon is the depth below which seawater is corrosive with respect to the calcium carbonate mineral aragonite. They show that although ensemble members generally agree on whether the saturation horizon shoals, there are differences of 30 or more years between ensemble members on the projected timing of local, abrupt shoaling of the aragonite saturation horizon.

Krumhardt et al. (2017) use the physical and biogeochemical projections of the CESM-LE to assess the long-term effects of climate change on coccolithophore growth and calcification. They find that, although growth and calcification are stimulated by the projected temperature increases over the 21st century, additional effects from increasing *p*CO2 and decreasing nutrient availability lead to an overall reduction in growth and calcification. The spread within the ensemble is used to test the statistical significance of these changes.

LEs can be used to force offline ecosystem models through efforts such as the Fisheries and Marine Ecosystem Model Intercomparison Project (Fish-MIP; Tittensor et al. 2018). However, they are currently limited by coarse model resolution of approximately 100 km. To robustly assess the regional influence of external forcing and internal variability on key coastal ecosystems, dynamical downscaling or higher resolution LE runs are required.

Outlook

The diverse applications of LEs in ocean biogeochemistry described herein underscore the multifaceted utility and transformative nature of this modeling tool. These applications have recently expanded beyond the initial scope of ToE studies, to the isolation of forced and unforced variability, and to ecosystem impacts and extreme events. Recent studies have used LEs for Observing System Simulation Experiments (e.g., Majkut et al. 2014; Gleoge et al. 2020), in generating observational LEs for biogeochemistry (Elsworth et al. 2020), in assessing decadal predictability through repeated ensemble initializations (Séférian et al. 2014; Li et al. 2016; Séférian et al. 2018; Lovenduski et al. 2019; Li et al. 2019; Brady et al. 2020; Fröhlicher et al. 2020; Krumhardt et al. 2020; Spring and Ilyina 2020), and in development of LE methodologies for partitioning sources of uncertainty for projections of marine biogeochemistry (Frölicher et al., 2016; Lovenduski et al., 2016; Schlunegger et al., 2020). LEs also provide a robust framework to isolate the biogeochemical signature of anthropogenic phenomena beyond greenhouse gas forcing and present a potentially powerful tool for exploring the effects of geoengineering (Lauvset et al. 2017) and nuclear war effects on marine ecosystems and biogeochemistry (Lovenduski et al. 2020).

Large potential remains for new insights into ocean biogeochemistry using the LE framework. While much of the initial LE work has focused on emergence and assessments of the long-term forced response, more work is needed on isolating the internal and external variability during the observed period, which showcases substantial variability on seasonal to multidecadal timescales. The forced component that emerges in the LE mean typically combines the effects of anthropogenic aerosols, volcanic aerosols, land-use change, and anthropogenic greenhouse gas radiative forcing. New "single forcing" experiments may be especially useful to isolate more precisely the biogeochemical influence of these different drivers. Additionally, future LE experimental design should consider the addition of external biogeochemical forcing by volcanoes through simulating atmospheric deposition of micronutrients (Hamme et al. 2010) and the use of prognostic (emissions-based), rather than diagnostic (concentration-based) CO2 (Arora et al. 2013).

Finally, much of the LE work has focused on individual model analysis. New opportunities exist for comparing biogeochemical responses across models through the US CLIVAR Working Group on Large Ensemble's Multi-Model Large Ensemble Archive (MMLEA, Deser et al. 2020), with a parallel effort focusing on the ocean carbon cycle (Schlunegger et al. 2020) as it differs across models. This perspective allows one to robustly

separate the effects of external forcing, model structure, and internal variability on ocean biogeochemistry—an opportunity that has long been sought after and is the result of many years of collaboration and hundreds of millions of hours of computational expense.

References

Arora, V. K., and Coauthors, 2013: Carbon-concentration and carbon-climate feedbacks in CMIP5 Earth system models. *J. Climate*, **26**, 5289-5314, doi:10.1175/JCLI-D-12-00494.1.

Bacastow, R., and E. Maier-Reimer, 1990: Ocean-circulation model of the carbon cycle. *Climate Dyn*., **4**, 95–125, doi:10.1007/BF00208905.

Bates, N. R., S. B. Moan, D. A. Hansell, and J. T. Mathis, 2006: An increasing CO2 sink in the Arctic Ocean due to sea-ice loss. *Geophys. Res. Lett.*, **33**, doi:10.1029/2006GL027028.

Bopp, L, and Coauthors, 2013: Multiple stressors of ocean ecosystems in the 21st century: Projections with CMIP5 models. *Biogeosciences*, **10**, 6225-6245, doi:10.5194/bg-10-6225-2013.

Brady, R. X., N. S. Lovenduski, M. A. Alexander, M. Jacox, and N. Gruber, 2019: On the role of climate modes in modulating the air–sea CO2 fluxes in eastern boundary upwelling systems. *Biogeosciences*, **16**, 329-346, doi:10.5194/bg-16-329-2019.

Brady, R. X., N. S. Lovenduski, S. G. Yeager, M. C. Long, and K. Lindsay, 2020: Skillful multiyear projections of ocean acidification in the California Current System. *Nat. Commun*., **11**, 1-9, doi:10.1038/s41467-020-15722-x.

Burger, F. A., T. L. Frölicher, and J. G. John, 2020: Increase in ocean acidity variability and extremes under increasing atmospheric CO2. *Biogeosci. Discuss*., doi:10.5194/bg-2020-22.

Cabré, A., I. Marinov, R. Bernardello, and D. Bianchi, 2015: Oxygen minimum zones in the Tropical Pacific across CMIP5 models: Mean state differences and climate change trends. *Biogeosciences*, **12**, 5429–5454, doi:10.5194/bg-12-5429-2015.

Carter, B. R., T. L. Frölicher, J. P. Dunne, K. B. Rodgers, R. D. Slater, and J. L. Sarmiento, 2016: When can ocean acidification impacts be detected from decadal alkalinity measurements. *Global Biogeochem. Cycles*, **30**, 595-612, doi:10.1002/2015GB005308.

Cheung, W. L., and T. L. Frölicher, 2020: Marine heatwaves exacerbate climate change impacts for fisheries in the northeast Pacific. *Sci. Rep*., **10**, doi:10.1038/s41598-020-63650-z.

Deser, C., A. Phillips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn*., **38**, 527-546, doi:10.1007/s00382-010-0977-x.

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat Climate Change*, **10**, 277–286, doi:10.1038/s41558-020-0731-2.

Di Lorenzo, E., and Coauthors, 2008: North Pacific Gyre Oscillation links ocean climate and ecosystem change. *Geophys. Res. Lett.*, **35**, doi:10.1029/2007GL032838.

Doney, S. C., V. J. Fabry, R. A. Feely, and J. A. Kleypas, 2009: Ocean acidification: The other CO2 problem. *Annu. Rev. Mar. Sci.*, **1**, 169-192, doi:10.1146/annurev.marine.010908.163834.

Eddebbar, Y. A., M. C. Long, L. Resplandy, C. Rödenbeck, K. B. Rodgers, M. Manizza, and R. F. Keeling, 2017: Impacts of ENSO on air-sea oxygen exchange: Observations and mechanisms. *Global Biogeochem. Cycles*, **31**, 901-921, doi:10.1002/2017GB005630.

Eddebbar, Y. A., K. B. Rodgers, M. C. Long, A. C. Subramanian, S.-P. Xie, and R. F. Keeling, 2019: El Niño–like physical and biogeochemical ocean response to tropical eruptions. *J. Climate*, **32**, 2627-2649, doi:10.1175/JCLI-D-18-0458.1.

Elsworth, G. W., N. S. Lovenduski, K. A. McKinnon, K. M. Krumhardt, and R. X. Brady, 2020: Finding the fingerprint of anthropogenic climate change in marine phytoplankton abundance. *Curr. Climate Change Rep*., **6**, 37–46, doi:10.1007/s40641-020-00156-w.

Fassbender, A. J., K. B. Rodgers, H. I. Palevsky, and C. L. Sabine, 2018: Seasonal asymmetry in the evolution of surface ocean PCO2 and PH thermodynamic drivers and the influence on sea-air CO2 flux. *Global Biogeochem. Cycles*, **32**, 1476–1497, doi:10.1029/2017GB005855.

Freeman, N. M., N. S. Lovenduski, D. R. Munro, K. M. Krumhardt, K. Lindsay, M. C. Long, and M. Maclennan, 2018: The variable and changing Southern Ocean silicate front: Insights from the CESM Large Ensemble. *Global Biogeochem. Cycles*, **32**, 752-768, doi:10.1029/2017GB005816.

Friedlingstein, P., M. Meinshausen, V. K. Arora, C. D. Jones, A. Anav, S. K. Liddicoat, and R. Knutti, 2014: Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *J. Climate,* **27**, 511–526, doi:10.1175/JCLI-D-12-00579.1.

Frölicher, T. L., J. L. Sarmiento, D. J. Paynter, J. P. Dunne, J. P. Krasting, and M. Winton, 2015: Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models. *J. Climate*, **28**, 862–886, doi:10.1175/JCLI-D-14-00117.1.

Frölicher, T. L., F. Joos, G.-K. Plattner, M. Steinacher, and S. C. Doney, 2009: Natural variability and anthropogenic trends in oceanic oxygen in a coupled carbon cycle–climate model ensemble. *Global Biogeochem. Cycles*, **23**, doi:10.1029/2008GB003316.

Frölicher, T. L., K. B. Rodgers, C. A. Stock, and W. L. Cheung, 2016: Sources of uncertainties in 21st century projections of potential ocean ecosystem stressors. *Global Biogeochem. Cycles*, **30**, doi:10.1002/2015GB005338.

Frölicher, T. L., L. Ramseyer, C. C. Raible, K. B. Rodgers, and J. Dunne, 2020: Potential predictability of marine ecosystem drivers. *Biogeosciences*, **17**, 2061–2083, doi:10.5194/bg-17-2061-2020.

Gloege, L., and Coauthors, 2020: Quantifying errors in observationally-based estimates of ocean carbon sink variability. *Earth Space Sci. Open Arch*., **24**, doi:10.1002/essoar.10502036.2.

Gruber, N., and Coauthors, 2019: The oceanic sink for anthropogenic CO2 from 1994 to 2007. *Science*, **363**, 1193-1199, doi:10.1126/science.aau5153.

Gruber, N., 2011: Warming up, turning sour, losing breath: ocean biogeochemistry under global change. *Philos. Trans. Roy. Soc. A*, **369**, 1980-1996, doi:10.1098/rsta.2011.0003.

Hamme, R. C., and Coauthors, 2010: Volcanic ash fuels anomalous plankton bloom in subarctic northeast pacific. *Geophys. Res. Lett*., **37**, doi:10.1029/2010GL044629.

Hawkins, E., and R. Sutton, 2011: The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dyn*., **37**, 407-418, doi:10.1007/s00382-010-0810-6.

Henson, S. A., C. Beaulieu, T. Ilyina, J. G. John, M. Long, R. Séférian, J. Tjiputra, and J. L. Sarmiento, 2017:Rapid emergence of climate change in environmental drivers of marine ecosystems. *Nat. Comm*., **8**, doi:10.1038/ncomms14682.

Ito, T., S. Minobe, M. C. Long, and C. Deutsch, 2017: Upper ocean O2 trends: 1958–2015. *Geophys. Res. Lett*., **44**, 4214–4223, doi:10.1002/2017GL073613.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.,* **96**, 1333-1349, doi:10.1175/BAMS-D-13-00255.1.

Keeling, R. F., A. Körtzinger, and N. Gruber, 2009: Ocean deoxygenation in a warming world. *Annu. Rev. Mar. Sci.*, **2**, 199–229, doi:10.1146/annurev.marine.010908.163855.

Kosaka, Y., and S.-P. Xie, 2013: Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature*, **501**, 403–407, doi:10.1038/nature12534.

Krumhardt, K. M., N. S. Lovenduski, M. D. Iglesias-Rodriguez, and J. A. Kleypas, 2017: Coccolithophore growth and calcification in a changing ocean. *Prog. Oceanogr*., **159**, 276-295, doi:10.1016/j.pocean.2017.10.007.

Krumhardt, K. M,, N. S. Lovenduski, M. C. Long, J. Y. Luo, K. Lindsay, S. Yeager, and C. Harrison, 2020: Potential predictability of net primary production in the ocean." *Global Biogeochem. Cycles*, **34**, doi:10.1029/2020GB006531.

Landschützer, P., and Coauthors, 2015: The reinvigoration of the Southern Ocean carbon sink. *Science*, **349**, 1221-1224, doi:10.1126/science.aab2620.

Landschützer, P., N. Gruber, D. C. E. Bakker, I. Stemmler, and K. D. Six, 2018: Strengthening seasonal marine CO2 variations due to increasing atmospheric CO2. *Nat. Climate Change*, **8,** 146–150, doi:10.1038/s41558-017-0057-x.

Lauvset, S. K., J. Tjiputra, and H. Muri, 2017: Climate engineering and the ocean: Effects on biogeochemistry and primary production. *Biogeosciences*, **14**, 5675–5691, doi:10.5194/bg-14-5675-2017.

Li, H., T. Ilyina, W. A. Müller, and P. Landschützer, 2019: Predicting the variable ocean carbon sink. *Sci. Adv.*, **5**, doi:10.1126/sciadv.aav6471.

Li, H., and T. Ilyina, 2018: Current and future decadal trends in the oceanic carbon uptake are dominated by internal variability. *Geophys. Res. Lett.*, **45,** 916–925, doi:10.1002/2017GL075370.

Li, H., T. Ilyina, W. A. Müller, and F. Sienz, 2016: Decadal predictions of the North Atlantic CO2 uptake. *Nat. Comm.*, **7**, doi:10.1038/ncomms11076.

Long, M. C., C. Deutsch, and T. Ito, 2016: Finding forced trends in oceanic oxygen. *Global Biogeochem. Cycles*, **30,** 381–397, doi:10.1002/2015GB005310.

Long, M. C., K. Lindsay, S. Peacock, J. K. Moore, and S. C. Doney, 2013: Twentieth-century oceanic carbon uptake and storage in CESM1 (BGC). *J. Climate*, **26**, 6775-6800, doi:10.1175/JCLI-D-12-00184.1.

Lovenduski, N. S., C. S. Harrison, H. Olivarez, C. G. Bardeen, O. B. Toon, J. Coupe, A. Robock, T. Rohr, and S. Stevenson, 2020: The potential impact of nuclear conflict on ocean acidification. *Geophys. Res. Lett.*, **47**, doi:10.1029/2019GL086246.

Lovenduski, N. S., G. A. McKinley, A. R. Fay, K. Lindsay, and M. C. Long, 2016: Partitioning uncertainty in ocean carbon uptake projections: Internal variability, emission scenario, and model structure. *Global Biogeochem Cycles,* **30,** 1276–87, doi:10.1002/2016GB005426.

Lovenduski, N. S., N. Gruber, S. C. Doney, and I. D. Lima, 2007: Enhanced CO2 outgassing in the Southern Ocean from a positive phase of the Southern Annular Mode. *Global Biogeochem Cycles*, **21**, doi:10.1029/2006GB002900.

Lovenduski, N. S, S. G. Yeager, K. Lindsay, and M. C. Lon, 2019: Predicting near-term variability in ocean carbon uptake. *Earth Syst. Dyn*., **10**, 45–57, doi:10.5194/esd-10-45-2019.

Maier-Reimer, E., and K. Hasselmann, 1987: Transport and storage of CO2 in the ocean ——an inorganic ocean-circulation carbon cycle model. *Climate Dyn*., **2**, 63–90, doi:10.1007/BF01054491.

Majkut, J. D., J. L. Sarmiento, and K. B. Rodgers, 201: A growing oceanic carbon uptake: Results from an inversion study of surface PCO2 data. *Global Biogeochem. Cycles*, **28**, 335–5, doi:10.1002/2013GB004585.

McKinley, G. A., A. R. Fay, N. S. Lovenduski, and D. J. Pilcher, 2017: Natural variability and anthropogenic trends in the ocean carbon sink. *Annu. Rev. Mar. Sci*., **9**, 125–150, doi:10.1146/annurev-marine-010816-060529.

McKinley, G. A., D. J. Pilcher, A. R. Fay, K. Lindsay, M. C. Long, and N. S. Lovenduski, 201: Timescales for detection of trends in the ocean carbon sink. *Nature*, **530**, 469–472, doi:10.1038/nature16958.

McKinley, G. A., M. J. Follows, J. Marshall, and S.-M. Fan, 2003: Interannual variability of air-sea O2 fluxes and the determination of CO2 sinks using atmospheric O2/N2. *Geophys. Res. Lett.*, **30**, doi:10.1029/2002GL016044.

Negrete-García, G., N. S. Lovenduski, C. Hauri, K. M. Krumhardt, and S. K. Lauvset, 2019: Sudden emergence of a shallow aragonite saturation horizon in the Southern Ocean. *Nat. Climate Change*, **9**, 313-317, doi:10.1038/s41558-019-0418-8.

Oeschger, H., U. Siegenthaler, U. Schotterer, and A. Gugelmann, 1975: A box diffusion model to study the carbon dioxide exchange in nature. *Tellus*, **27**, 168–192, doi:10.3402/tellusa.v27i2.9900.

Rodgers, K. B., J. Lin, and T. L. Frölicher, 2015: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences*, **12**, 3301–3320, doi:10.5194/bg-12-3301-2015.

Sarmiento, J. L., J. C. Orr, and U. Siegenthaler, 1992: A perturbation simulation of CO2 uptake in an ocean general circulation model. *J. Geophys. Res*., **97**, 3621–3645, doi:10.1029/91JC02849.

Schlunegger, S., K. B. Rodgers, J. L. Sarmiento, T. L. Frölicher, J. P. Dunne, M. Ishii, and R. Slater, 2019: Emergence of anthropogenic signals in the ocean carbon cycle. *Nat. Climate Change*, **9,** 719–725, doi:10.1038/s41558-019-0553-2.

Schlunegger S., K. Rodgers, J. L. Sarmiento, T. Ilyina, J. P. Dunne, T., Takano, M. C. Long, T. L. Frölicher, R. Slater, and F. Lehner, 2020: Time of emergence & large ensemble intercomparison for ocean biogeochemical trends. *Global Biogeochem. Cycles*, Accepted.

Séférian, R., L. Bopp, M. Gehlen, D. Swingedouw, J. Mignot, E. Guilyardi, and J. Servonnat, 2014: Multiyear predictability of tropical marine productivity *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 11646–11651, doi:10.1073/pnas.1315855111.

Séférian, R., S. Berthet, and M. Chevallier, 2018: Assessing the decadal predictability of land and ocean carbon uptake. *Geophys. Res. Lett*., **45**, 2455–2466, doi:10.1002/2017GL076092.

Siegenthaler, U., 1983: Uptake of excess CO2 by an outcrop-diffusion model of the ocean. *J. Geophys. Res. Ocean*., **88**, 3599–3608, doi:10.1029/JC088iC06p03599.

Spring, A., and T. Ilyina, 2020: Predictability horizons in the global carbon cycle inferred from a perfect-model framework. *Geophys. Res. Lett.*, **47**, doi:10.1029/2019GL085311.

Stevenson, S., A. Capotondi, J. Fasullo, and B. Otto-Bliesner, 2019: Forced changes to twentieth century ENSO diversity in a last Millennium context. *Climate Dyn*., **52**, 7359-7374, doi:10.1007/s00382-017-3573-5.

Takahashi, T., and Coauthors, 2009: Climatological mean and decadal change in surface ocean pCO2, and net sea-air CO2 flux over the global oceans. *Deep Sea Res., Part* II, **56**, 554–577, doi:10.1016/j.dsr2.2008.12.009.

Tans, P. P., I. Y. Fung, and T. Takahashi, 1990: Observational contrains on the global atmospheric CO2 budget. *Science*, **247**, 1431–1438, doi:10.1126/science.247.4949.1431.

Timmermann, A., J. Oberhuber, A. Bacher, M. Esch, M. Latif, and E. Roeckner, 1999: Increased El Niño frequency in a climate model forced by future greenhouse warming. *Nature*, **398**, 694-697, doi:10.1038/19505.

Tittensor, D. P., and Coauthors, 2018: A protocol for the intercomparison of marine fishery and ecosystem models: Fish-MIP v1.0. *Geosci. Model Dev.*, **11**, 1421–1442, doi:10.5194/gmd-11-1421-2018.

Verdy, A., S. Dutkiewicz, M. J. Follows, J. Marshall, and A. Czaja, 2007: Carbon dioxide and oxygen fluxes in the Southern Ocean: Mechanisms of interannual variability. *Global Biogeochem. Cycles*, **21**, doi:10.1029/2006GB002916.
Wanninkhof, R., and Coauthors, 2013: Global ocean carbon uptake: Magnitude, variability and trends. *Biogeosciences*, **10**, 1983–2000, doi:10.5194/bg-10-1983-2013.
Woods, J. D., 1985: The World Ocean Circulation Experiment. *Nature*, **314**, 501–511, doi:10.1038/314501a0.
Yin, J. H., 2005: A consistent poleward shift of the storm tracks in simulations of 21st century climate. *Geophys. Res. Lett.*, **32**, doi:10.1029/2005GL023684.
Zanchettin, D., C. Timmreck, H.-F. Graf, A. Rubino, S. Lorenz, K. Lohmann, K. Krüger, and J. Jungclaus, 2012: Bi-decadal variability excited in the coupled ocean– atmosphere system by strong tropical volcanic eruptions. *Climate Dyn.*, **39**, 419–444, doi:10.1007/s00382-011-1167-1.

Evaluating the internal variability and forced response in Large Ensembles

Laura Suarez-Gutierrez, Nicola Maher, and Sebastian Milinski

Max-Planck-Institut für Meteorologie, Germany

Surface temperatures and all variables in the climate system fluctuate around their long-term evolving forced state due to the chaotic effect of internal variability. Real-world observations offer only one amongst many possible combinations of these fluctuations, making it difficult to distinguish the effect of internal variability from the forced response to external drivers. In contrast, initial-condition large ensembles (LEs) consist of up to hundreds of simulations of a single climate model under the same time-evolving external forcing conditions, which differ only due to the effect of chaotic internal variability. This means that when large enough LEs allow a precise quantification of both the time-evolving forced response, represented by the ensemble mean, and the internal variability, represented by the spread of possible fluctuations around this mean.

Due to their design, LEs allow for a more effective climate model evaluation. We can use LEs to determine whether observations fall within the ensemble spread simulated by each model. We exploit this potential of LEs to evaluate how well climate models capture the internal variability and forced response in observations, without the need to separate both quantities in the observations, by applying a methodological evaluation framework based on probabilistic forecast verification (Hamill 2001; Suarez-Gutierrez et al. 2018; Maher et al. 2019). This evaluation framework allows us to determine model performance more robustly than before, by assessing whether current climate models capture the long-term trajectory of the climate system as well as the possible range of fluctuations around this trajectory caused by internal variability in any given region and time period.

Here, we use this framework to evaluate historical near-surface air temperatures over North America in LEs from six comprehensive fully-coupled climate models in the Multi-Model Large Ensemble Archive (MMLEA; Deser et al. 2020) provided by the US CLIVAR Working Group on LEs: CanESM2, CESM-LE, CSIRO-MK3.6,GFDL-CM3, GFDL-ESM2M, MPI-GE; as well as in the Observational LE (OBS-LE). In contrast to the six model LEs, OBS-LE is a statistical product that combines the simulated time-evolving forced response from CESM-LE with a synthetic statistical estimate of internal variability derived from observations from the Berkeley Earth Surface Temperature (BEST) dataset for temperature (McKinnon and Deser 2018).

Results

Time series and rank histogram analysis

For the hypothetical case of an LE that perfectly represents the combined effect of the real-world forced response and internal variability, a sufficiently long sample of observations should fall across all of the ensemble spread with no preferred frequency, and mainly occur within the ensemble maximum and minimum limits. We evaluate this by computing time series and rank histograms of annually averaged North American near-surface air temperature anomalies (SAT) with respect to the reference period of 1961–1990 compared to CRUTEM4 observations (Figure 1).

The time series in Figure 1 show the ensemble maxima and minima as well as the 75th percentile central ensemble range (i.e., 12.5th to 87.5th percentile range), together with observations. The rank histograms shown in Figure 1 represent the frequency with which observations take each place in a list of ensemble members ordered by ascending SAT values for each year (Hamill 2001). The rank is zero if the observed SAT for a given year is lower than each SAT simulated by all the ensemble members for that year. If the observed SAT is higher than all simulated SATs, the rank is n, the number of ensemble members. For a long enough observational record that is adequately simulated, observations should occur in all ranks with uniform frequency, thus resulting in a flat rank histogram. In contrast, a non-flat rank histogram indicates a model bias in either the variability or forced response. This is the case for CanESM2, CSIRO-MK3.6, and GFDL-CM3, which show sloped rank histograms with disproportionately large low-rank frequencies. Thus, these ensembles overestimate the historical forced warming compared to observations. Observations occur frequently in the lower half of these ensembles, or below the ensemble minima, either during the entire observational record as for GFDL-CM3, or only in recent decades or early historical period, as for CanESM2 and CSIRO-MK3.6 respectively. The remaining LEs — CESM-LE, GFDL-ESM2M, MPI-GE, and OBS-LE — show relatively flat rank histograms. This indicates that these LEs cover the time-evolving observational spread in North American SATs adequately, with observations occurring uniformly across the ensemble spreads and mostly within the ensemble limits.

The LEs with longer simulation lengths, CESM-LE and in particular CSIRO-MK3.6 and MPI-GE, also appear to have larger SAT variability in the 19th and early 20th Centuries than in recent decades. This variability, represented by the ensemble spread of SAT, decreases in recent decades to maximum to minimum annual SAT ranges of around 1.5 to 2.0 °C, of similar magnitude across all LEs. We also find year-to-year variability in the ensemble maxima and minima SAT larger than 0.5 °C across all six climate models LEs.

By contrast, and due to its experimental design, OBS-LE shows substantially less year-to-year variability in the ensemble maxima and minima. This could arise from the large ensemble size of 1,000 members resulting in the saturation of the SAT ensemble spread on yearly timescales. However, this year-to-year variability remains comparatively low when only the first 100 members of OBS-LE are considered, and is also lower than the variability that we could expect from normally distributed data (not shown), indicating a potential under-sampling of the distribution tails. This suggests

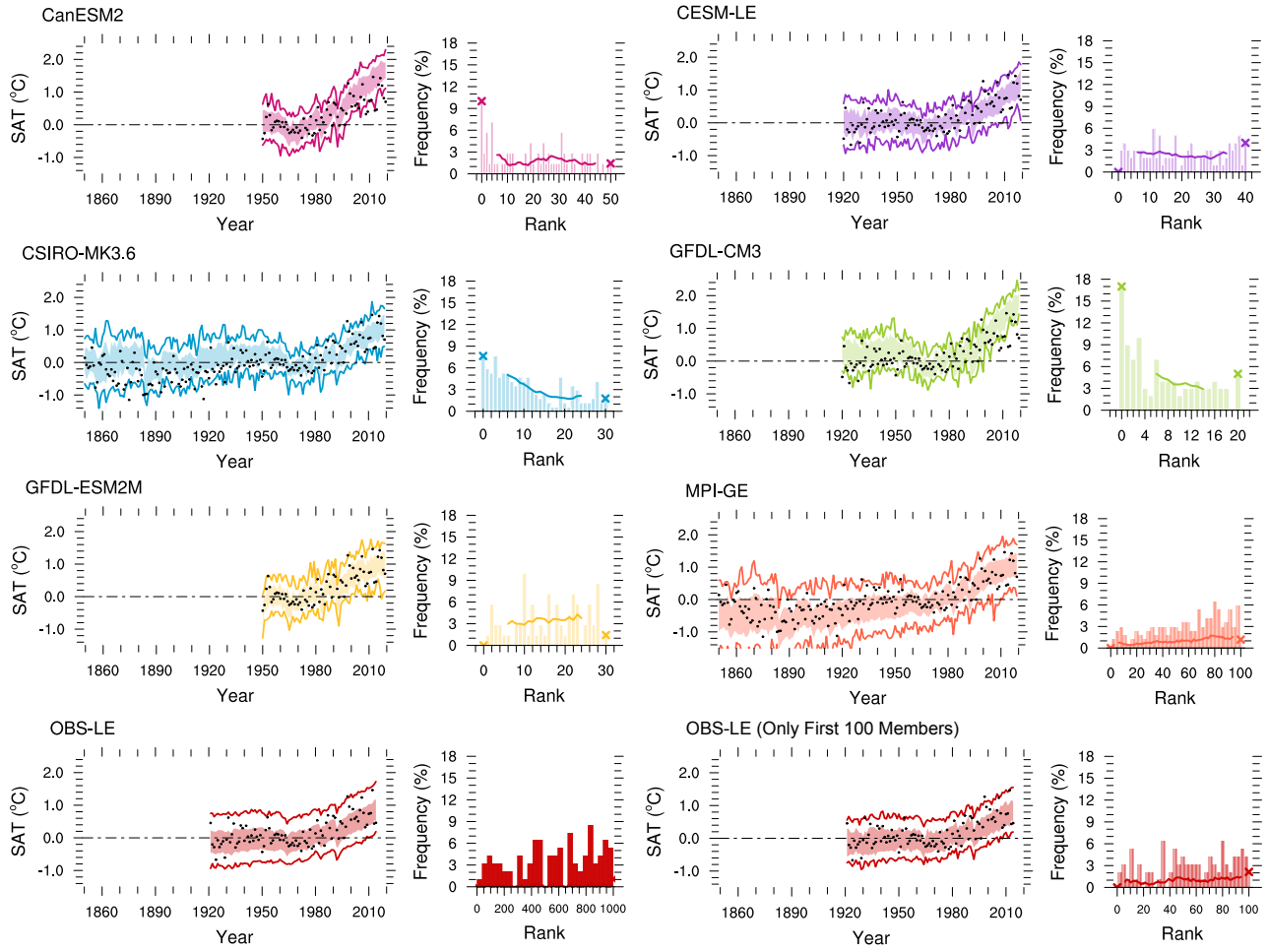


Figure 1: Time series and rank histograms of annual SAT over North America. Time series of annual land-surface SAT anomalies simulated by each LE (colored) and CRUTEM4 observed anomalies (black circles) for the period 1850–2019 (left column). Lines represent ensemble maxima and minima, shading represents the central ensemble range within the 75th percentile (12.5th to 87.5th percentiles). Rank histograms show the frequency of each place that CRUTEM4 observations would take in a list of ensemble members ordered by ascending SAT values (right column). Crosses represent the frequency of minimum (0) and maximum (number of members; n) ranks; lines illustrate the histogram's slope as the moving 10-rank mean. Frequencies are normalized to percentage. Bin sizes are 1 rank, except for MPI-GE and OBS-LE where bin sizes from ranks 1 to n-1 are 3 and 37 ranks, respectively, to aid visualization. Anomalies are relative to the period 1961–1990. Temperature anomalies are averaged over land-surface grid cells where observations are available in the [17.5–52.5°N, 62.5°W–127.5°W] domain.

that OBS-LE may underestimate the intensity of the most extreme SAT events. This could result from the lack of sufficiently large samples of observed low-probability events, due to relatively short observational record, which leads to not only the potential underestimation of the intensity of extreme events but also complicates the robust estimation of their likelihood.

Although this comparatively low year-to-year variability might indicate that OBS-LE underestimates the intensity of low-probability events at the tails of the ensemble distribution, OBS-LE offers the most adequate representation of the combination of the internal variability and forced response in observed SAT over North America throughout the historical observational record. The climate model LEs that capture both quantities in observations most adequately are GFDL-ESM2M, CESM-LE and MPI-GE.

Spatial representation of the combined forced response and internal variability in observations

Based on the concepts in the previous section, we now evaluate how different LEs capture the internal variability and forced response in observations at the grid-cell level by identifying three different possible biases (Figure 2). First, we evaluate how often observations lie either below or above the ensemble limits in each grid cell. We distinguish between regions where 5% or more of the time observations fall below the ensemble minimum (blue shading) and above the ensemble maximum (red shading). If only one of these biases occurs in a region, the model respectively over- or under-estimates the forced response in observations. Alternatively, such a bias could also be caused by a bias in the skewness of the probability distribution for non-normally distributed variables. If both of these biases occur at the same location, this means that observations fall below and also above the ensemble limits, either over the entire period of analysis (indicating the model does not sufficiently capture the observed variability) or during specific periods (indicating a likely change in the sign of the model bias over time).

The third metric of model performance highlights regions where observations cluster more than expected within the central 75th percentile range of the simulated ensembles. For the ideal case in which observations are uniformly distributed across the ensemble and exhibit a flat rank histogram, observed values would lie within the central 75th percentile ensemble range (12.5th–87.5th percentiles) around 75% of the time. Here we identify areas where observations occur in the central ensemble range more than 80% of the time (gray shading in Figure 2), indicating that the model overestimates internal variability. This bias results in simulated extreme events at the tails of the ensemble distribution that are systematically more intense than observed. Note that this type of bias can only be robustly identified when the simulated distribution adequately captures the forced response in observations, and when evaluated over a period long enough to sufficiently sample the timescales of internal variability under study.

White areas without any shading in Figure 2 indicate that none of the three biases occurs to a substantial degree, indicating that the ensembles simulate a time-evolving forced response and range of variability around this response that are comparable to those in observations for the whole length of their simulations. Thus, in these areas, our evaluation framework indicates that the models adequately capture the forced response and internal variability in observed surface temperatures. The percentage of white areas over North America represents areas with no substantial biases for each LEs (upper right corners in Figure 2), and indicates that OBS-LE, with 85.9%, offers the most adequate spatial representation of the combined internal variability and forced response in observed historical SAT. MPI-GE, with 46.6% of white areas, offers the best representation of historical SAT over North America amongst the model LEs, followed by CanESM2, CESM-LE, and GFDL-ESM2M.

The predominance of blue shading over red shading in Figure 2 for CanESM2, and especially CSIRO-MK3.6 and GFDL-CM3, indicates that observations fall below the ensemble minima more frequently and over larger regions than they fall above the ensemble maxima for these models. These are the same models that show overestimated forced warming compared to observations in Figure 1. Observations exceed the ensemble maxima over the East Coast and Gulf of Mexico area for CESM-LE and CSIRO-MK3.6 (red shading in Figure 2b and c), indicating that these ensembles underestimate the intensity of warm near-surface air temperature extremes in these areas.

Over the Caribbean Islands and the Baja California Peninsula, observations occur both below and above ensemble limits with high frequency (overlapping blue and red shading in Figure 2). This indicates that models underestimate observed SAT variability in these

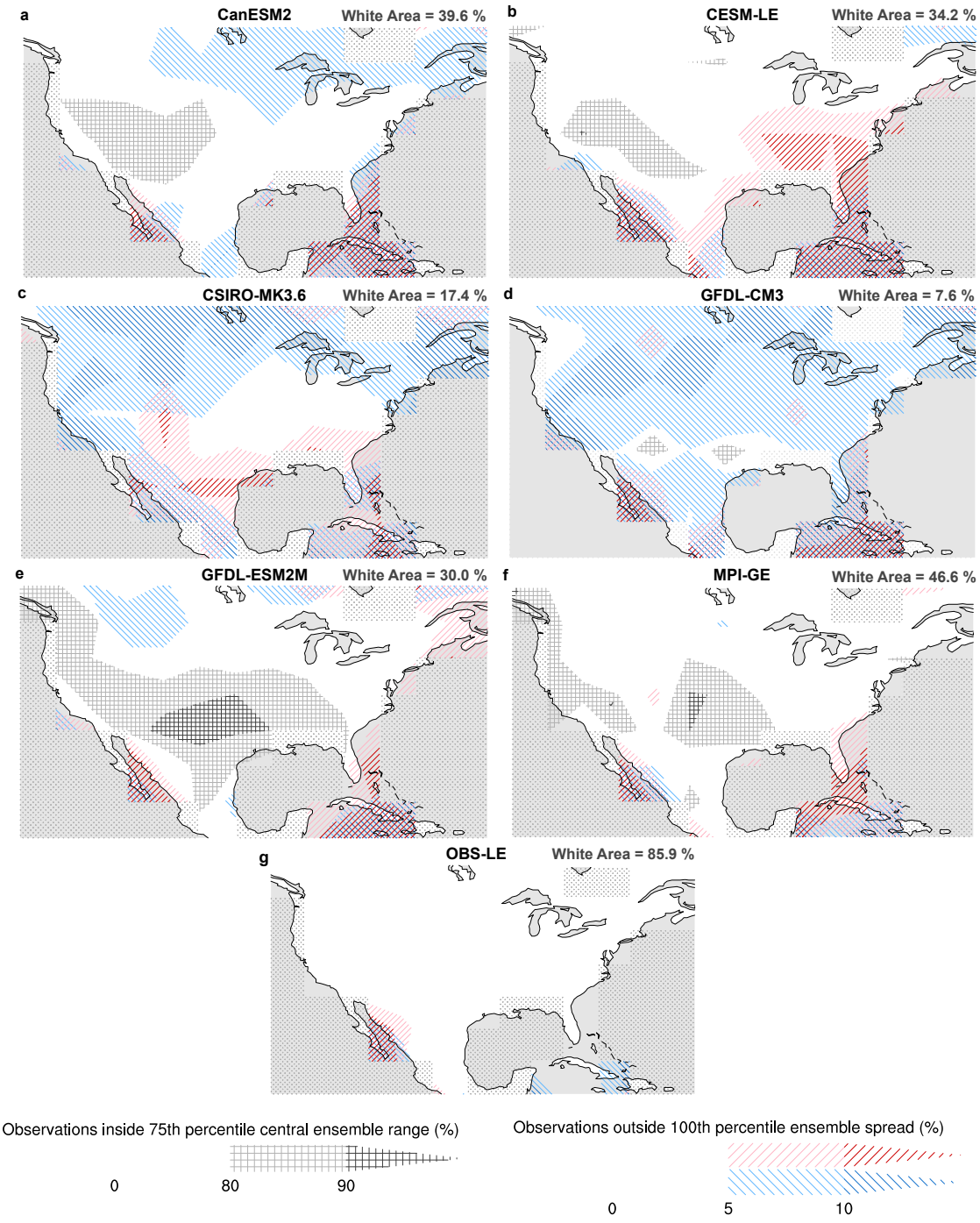


Figure 2: Evaluation of internal variability and forced response in annual SATs. Evaluation of annual SAT anomalies simulated by different LEs compared to CRUTEM4 observed anomalies from 1850, or each LE starting year, until 2019. Red shading represents the percentage of time that the observed yearly anomaly is larger than the ensemble maximum; blue shading represents the percentage of time that the observed yearly anomaly is lower than the ensemble minimum. Gray hatching represents how often observations cluster within the 75th percentile bounds of the ensembles (12.5th to 87.5th percentiles). Dotted areas are excluded from our analysis due to CRUTEM4 observations being available for less than 10 years. Percentages of white area in the upper right corners represent the percentual area of North America where none of these biases occur to a substantial degree for each LEs. Anomalies are relative to the period 1961–1990. Model output data are regridded to match the observational grid.

regions, likely due to the effect of model resolution and complex orography in confounding land versus ocean in these grid cells. Lastly, observations cluster in the central ensemble ranges of several models, including CanESM2, CESM-LE, MPI-GE, and especially GFDL-ESM2M, over the West Coast and Central US, indicating that these models overestimate the variability in these regions (gray shading in Figure 2).

OBS-LE shows no substantial biases over North America, with the exception of the underestimation of SAT variability over the Baja California Peninsula (Figure 2g). Our results indicate that OBS-LE offers the most adequate spatial representation of the internal variability and forced response in observed historical SAT over this region. In agreement with the results in Mckinnon and Deser 2018 for 50-year trends, we find that OBS-LE shows only minor biases in annual SATs over most of the Northern Hemisphere; while it exhibits underestimated annual SAT variability compared to observations over large areas in the low latitudes (not shown). Over these regions, OBS-LE fails to cover the observed variability range in SATs, with observed extreme anomalies beyond the OBS-LE maximum and minimum values over more than 10% of the years. This could result from a combination of the comparatively lower variability at the tails of the OBS-LE distribution identified in Figure 1, that could be more prominent in these areas, as well as an increased spatial and temporal observational sparsity in these regions that could affect the statistical processing used to generate OBS-LE.

Comparison of internal variability

Following our evaluation of the forced response and internal variability in LEs, we can now determine which LEs provide the most realistic simulations of internal variability in annual SAT over North America to better estimate the internal variability in the real world. Here, we measure the magnitude of internal variability in the model LEs and OBS-LE as the 2.5th to 97.5th percentile ensemble spread averaged over the period 1950–1990 (Figure 3a-g). We restrict this analysis to the period 1950–1990 to ensure contributions from all LEs and to minimize the

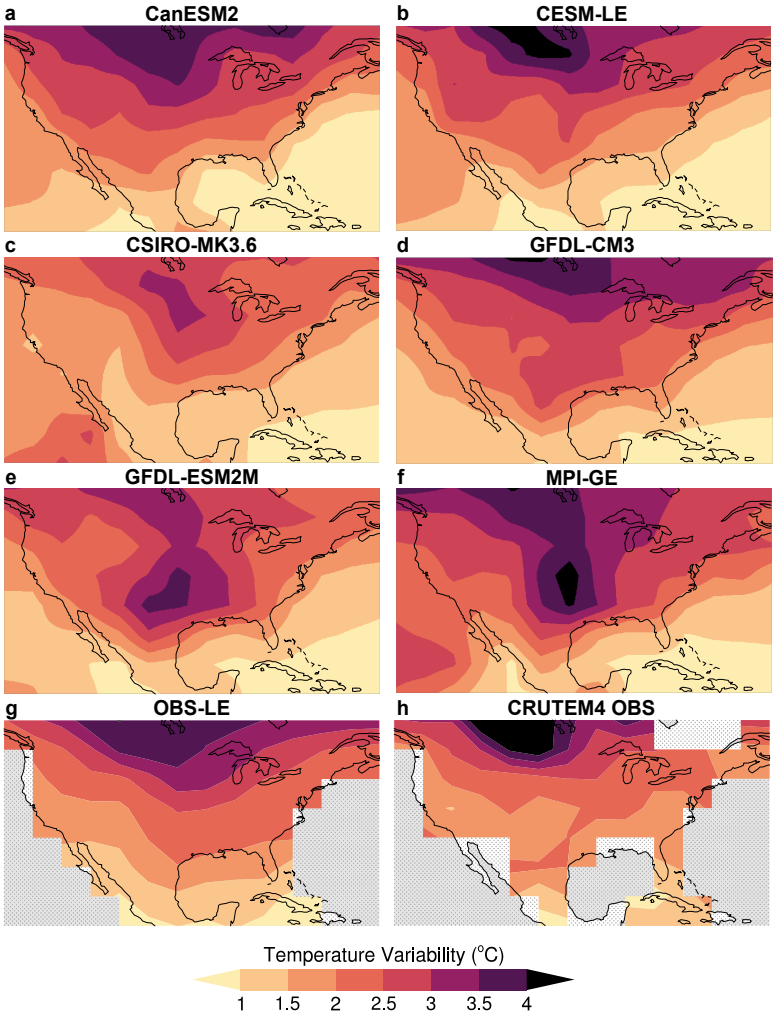


Figure 3: Variability in annual surface temperatures. (a-g) Ensemble spread for annual SAT anomalies by different model LEs and OBS-LE averaged for the period of 1950–1990 measured as the difference between the 2.5th to 97.5th annual SAT percentiles. (h) SAT spread in CRUTEM4 observations measured as the difference between the 2.5th to 97.5th percentiles of the whole distribution for the period of 1950–1990 of annual SAT anomalies. Simulated data are regridded to match the observational grid.

potential effect of the forced response. For comparison, we compute the range of internal variability in CRUTEM4 observations, estimated directly as the 2.5th to 97.5th percentile range in the distribution of non-detrended observed annual SAT anomalies in the same period (Figure 3h). The simulated amplitude of SAT internal variability ranges from 2 to 4°C over most of the continental land area across the different model LEs (Figure 3a-f) and OBS-LE (Figure 3g). The observational estimate (Figure 3h) is generally larger than the LE estimates, in particular over the northern part of the domain, and decreases more steeply with decreasing latitude. However, unlike the LEs estimates, the observational estimate of internal variability could be affected by the confounding effect of the forced response in observations.

OBS-LE, which by design most adequately captures the forced response and internal variability in SAT over North America, exhibits a stratified pattern with variability increasing polewards, which is not completely captured by any of the LEs. Two of the four LEs that most adequately capture North American SATs, GFDL-ESM2M and MPI-GE (Figure 3e and f), simulate hotspots of too high SAT variability over the central United States and Gulf of Mexico region, in agreement with the areas of overestimated variability in Figure 2. These hotspots exhibit SAT variability ranges of 3.5°C to more than 4.0°C, almost twice as large as the SAT variability in other LEs. This indicates that in these areas, these two ensembles simulate annual mean SAT extremes systematically more intense than those observed, possibly due to an overestimation of the cold tail of the distribution during the summer months (not shown).

Summary and conclusions

We use a novel framework exploiting the power of large ensembles to evaluate historical temperatures over North America in six comprehensive, fully-coupled climate models, as well as in the observational ensemble OBS-LE. This framework is based on a simple approach: evaluating whether observations occur evenly across the ensemble spread of simulations, and whether they occur mainly within the limits of this spread. Our evaluation shows that the experimental design in OBS-LE results in the most adequate representation of the combined effect of the forced response and internal variability in observed temperatures over North America. The climate model LEs that provide the best representation according to our metrics are MPI-GE, CanESM2, CESM-LE, and GFDL-ESM2M, suggesting that these LEs are the best choice for investigating future temperature projections over this region. Our evaluation framework highlights MPI-GE as the model LE that most adequately captures the combined forced response and internal variability in observed North American surface temperatures for the period 1850–2019, with the largest area with no substantial biases, 46.6% of the North American region.

Several models show similar biases over similar regions, such as an overestimation of temperature variability in Central North America and an underestimation of variability over the Caribbean Islands and the Baja California Peninsula, likely due the combination of model resolution and complex orography in these regions. Some models overestimate recent forced warming over North America beyond the range of plausible fluctuations caused by internal variability. Our results show that models do not consistently over- or underestimate internal variability in surface temperatures, and that models that perform adequately over one region will not necessarily do so in another.

Overall, this evaluation framework provides a new and more robust approach to determine model performance, allowing users to decide which models are most appropriate for their variable and region of interest, by highlighting which models offer the most adequate representation of the real-world internal variability and forced response.

Acknowledgments

We acknowledge the modeling groups that created the model data used here, as well as the US CLIVAR Working Group on Large Ensembles for facilitating

access to these data in the Multi-Model Large Ensemble Archive (Deser et al. 2020). We would also like to acknowledge the groups that developed and facilitated the observational compilations used here: the Climatic Research Unit (University of East Anglia) in conjunction with the Hadley Centre, UK Met Office (Jones et al. 2012).

Data and methods

We include LEs from six coupled climate models in the US CLIVAR MMLEA (Deser et al. 2020) as well as the synthetic product OBS-LE based on observations (Table 1). Each of the climate model LEs comprises several simulations for one fully coupled climate model that differ only in their initial state, and evolve under one specific set of forcing conditions. However, the ensembles differ in their number of simulations, in how sensitive the model is to increasing CO2, or in the method used for the initialization of their members. When available, historical simulations are extended with one available future forcing scenario to cover the entire observational record. We also use surface temperature observations from the CRUTEM4 (Jones et al. 2012) dataset for comparison to the LE simulations. All simulated data are regridded to match the coarser resolution of CRUTEM4 observations and transformed to anomalies with respect to the 1960–1991 climatological period.

The methodological framework demonstrated in this paper was first used in Suarez-Gutierrez et al. 2018 to evaluate European summer temperature and precipitation in MPI-GE; and further expanded to evaluate global annual mean temperatures in Maher et al. 2019, and global summer maximum temperatures in Suarez-Gutierrez et al. 2020.

Table 1: Details of LE experiments analysed from the Multi-Model Large Ensemble Archive (Deser et al. 2020). Experiment name, number of members, simulated years used, forcing scenarios and references of LE experiments included. All experiments include historical forcing (Hist.) until 2005, except for OBS-LE, which is based on historical observed temperatures (see McKinnon and Deser, 2018). Historical simulations are extended beyond 2005 using one future forcing scenario.

LE Experiment	Members	Years	Forcing	Reference
CanESM2	50	1950-2018	Hist + RCP8.5	Kirchmeier-Young et al. 2017
CESM-LE	40	1920-2018	Hist + RCP8.5	Kay et al. 2015
CSIRO-MK3.6	30	1850-2018	Hist + RCP8.5	Jeffrey et al. 2013
GFDL-CM3	20	1920-2018	Hist + RCP8.5	Sun et al. 2018
GFDL-ESM2M	30	1950-2018	Hist + RCP8.5	Rodgers et al. 2015
MPI-GE	100	1850-2018	Hist + RCP8.5	Maher et al. 2019
OBS-LE	1000	1920-2014	Hist + RCP8.5	McKinnon et al. 2017

References

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277-286, doi:10.1038/s41558-020-0731-2.

Hamill, T. H., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Jeffrey S., L. Rotstayn, M. Collier, S. Dravitzki, C. Hamalainen, C. Moeseneder, K. Wong, and J. Syktus, 2013: Australia's CMIP5 submission using the CSIRO-Mk 3.6 model. *Aust. Meteor. Oceanogr. J.*, **63**, 1–13, doi:10.22499/2.6301.001.

Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice, 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *J. Geophys. Res.: Atmos.* **117**, doi:10.1029/2011JD017139.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.,* **96**, 1333–1349, doi:10.1175/BAMS-D-13-00255.1.

Kirchmeier-Young, M. C., F. W. Zwiers, and N. P. Gillett, 2017: Attribution of extreme events in Arctic Sea ice extent. *J. Climate,* **30**, 553–571, doi:10.1175/JCLI-D-16-0412.1.

McKinnon, K. A. and C. Deser, 2018. Internal Variability and Regional Climate Trends in an Observational Large Ensemble. *Journal of Climate* **31**, 6783–6802 doi:10.1175/JCLI-D-17-0901.1.

Maher, N., and Coauthors, 2019: The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst.,* **11**, 2050–2069, doi:10.1029/2019MS001639.

Rodgers, K. B., J. Lin, and T. L. Frölicher, 2015: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences,* **12**, 3301–3320, doi:10.5194/bg-12-3301-2015.

Suarez-Gutierrez, L., W. A. Müller, C. Li, and J. Marotzke, 2018: Internal variability in European summer temperatures at 1.5 °C and 2 °C of global warming. *Environ. Res. Lett.,* **44**, 5709–5719, doi:10.1088/1748-9326/aaba58.

Suarez-Gutierrez, L., W. A. Müller, C. Li, and J. Marotzke, 2020: Hotspots of extreme heat under global warming. *Climate Dyn.,* **55**, 429-447, doi:10.1007/s00382-020-05263-w.

Sun, L., M. Alexander, and C. Deser, 2018: Evolution of the global coupled climate response to Arctic Sea ice loss during 1990–2090 and its contribution to climate change. *J. Climate,* **31**, 7823–7843, doi:10.1175/JCLI-D-18-0134.1.

Detecting forced changes in internal variability using Large Ensembles: On the use of methods based on the "snapshot view"

Tímea Haszpra[1], Dániel Topál[2,3], Mátyás Herein[1]

[1]Eötvös Loránd University, Hungary
[2]Research Center for Astronomy and Earth Sciences, Hungary
[3]University of California-Santa Barbara, USA

Anthropogenic activities contribute to the rising level of greenhouse gas concentrations in the atmosphere at a rate of approximately 1% per year providing a time-dependent external radiative forcing on the climate system (Peters et al. 2020). Associated changes in several climate variables (e.g., global mean surface temperature) are thought to have emerged from unforced natural internal processes of the climate system allowed by the characteristics of the dissipative chaotic nature of the climate dynamics, i.e., from internal variability (Hawkins et al. 2020). In addition to tangible consequences of anthropogenic forcing affecting the climate system (e.g., the dramatic Arctic sea ice retreat (Screen and Simmonds 2010)), simultaneous, less apparent changes occurring on low-frequency timescales demand effort to deal with. These include changes in internal variability due to the non-stationary anthropogenic forcing, that represents additional uncertainty affecting future model projections on top of internal variability, scenario and model uncertainty (Hawkins and Sutton 2009; Deser et al. 2012; Wettstein and Deser 2014; Lehner et al. 2020).

Although previous studies using observations and multi-model single-member simulations successfully detected, for example, changes in the jet-stream variability (Barnes and Polvani 2013) or in the variance of Northern Hemisphere (NH) temperature (Screen 2014) due to anthropogenic forcing, traditional methods based on long-term temporal statistics unavoidably make use of discrete time windows subjectively chosen from a continuously time dependent system (i.e., the changing climate). In addition, separating the effects of model structural differences and internal variability in multi-model ensembles is challenging (Merrifield et al. 2019). State-of-the-art Single Model Initial-condition Large Ensemble (SMILE) simulations (Kay et al. 2015; Maher et al. 2019; Deser et al. 2020) – that account for the chaotic behavior of the climate system with perturbed initial condition runs of the same model – offer a way forward for new perspectives on externally-forced changes in internal variability. Here, we outline an approach for analyzing SMILEs called the snapshot view, which offers a mathematically exact and elegant formulation and

the potential to complement previous diagnostics with ensemble-based statistics.

Theoretical background – The snapshot view

The concept of the so-called "snapshot view" was introduced into dynamical system theory to understand how nonautonomous dynamics behave when subjected to general time dependent forcings. Romeiras et al. (1990) drew attention to an interesting feature of dissipative dynamical systems: the fact that a single long "noisy" trajectory traces out a fuzzy shape, while an ensemble of motions starting from many different initial conditions, using the same noise realization along each trajectory, creates a structured fractal pattern at any instant. This ensemble-related pattern, the snapshot chaotic attractor, continuously changes its shape, in contrast to traditional chaotic attractors, which are time-independent (Lorenz 1963; Ott 1993). The concept of snapshot attractors has been used to understand a variety of time-dependent physical phenomena (see e.g., Pikovsky 1984; Yu et al. 1990; Serquina et al. 2008; Ku et al. 2015; Vincze et al. 2017). However, it was not until Ghil et al. (2008) and Checkroun et al. (2011) pointed out its potential importance to the field of climate dynamics that it began to be more widely applied in climate science and that this concept (also called the "pullback attractor") was relevant for the understanding of anthropogenic climate change.

Deterministic (noise-free) snapshot attractors capture the essence of an unpredictable dynamical system under changing conditions (Bódai and Tél 2012; Pierini 2014; Drótos et al. 2015). The traditional way of obtaining a chaotic attractor by means of a "single trajectory" is not equivalent to the "ensemble" method (ergodicity does not hold in systems subjected to forcings of general time-dependence). One has to choose between the two approaches, and it is the ensemble-based snapshot view that is appropriate for a faithful statistical representation of the possible distribution of a given quantity at any time instant in a changing climate. The reason is that the ensemble also represents a natural probability distribution, supported by the snapshot attractor.

The basic features of the snapshot view valid for any dissipative system subjected to general forcing can be summarized as follows (Drótos et al. 2015):

- Conclusions based on single trajectories may be misleading since such trajectories are unpredictable, thus not representative.
- On the contrary, ensemble properties, including the natural probability distributions (which set in after the initial conditions are "forgotten" in a numerical simulation) are fully predictable in a statistical sense (in harmony with general properties of chaotic systems (Tél and Gruiz 2006)).
- An instantaneous characterization of the system becomes possible (as properly expressed by the adjective "snapshot"), and the use of (occasionally biased) temporal averages can fully be avoided.
- It offers a straightforward way to analyze internal variability in a changing system, e.g., by means of statistical quantifiers of the instantaneous probability distribution.
- Because the instantaneous (snapshot) statistics are available at each time instant, the forced changes in any quantity, such as the internal variability, can be determined by analyzing the time series of snapshot values by means of the traditional tools of time series analysis.

In the particular example of climate change, the snapshot view can equivalently be formulated as the theory of *parallel climate realizations* (Herein et al. 2017; Tél et al. 2019). Qualitatively speaking, one can imagine many copies of the Earth system moving on different dynamical paths, each being subjected to the same physical laws and forcings. As a generalization of Leith's observation (Leith 1978), parallel climate realizations constitute an ensemble of a large number of members, and the snapshot taken over the ensemble (the snapshot attractor) represents the plethora of permitted climate states at that instant.

Utilizing the snapshot view to detect forced changes in internal variability

In this section we reveal how the *snapshot view* allows for surprisingly simple practices to detect forced changes in internal variability via utilizing SMILE simulations. Here, we focus on modes of large-scale internal atmospheric circulation variability (so-called "teleconnection patterns"), which may change due to anthropogenic forcing. The question arises how to characterize changes in these modes as a result of climate change, since certain characteristics of the linkages between the teleconnection patterns and other climate variables, for example precipitation or air temperature, may also change even within a carefully chosen time window (Drótos et al. 2015; Herein et al. 2016; Herein et al. 2017; Tél et al. 2019). Therefore, we need to reconsider these methodologies when aiming to detect forced changes in internal variability.

Our previous research (Haszpra et al. 2020b) exemplified a novel means of analyzing changes in modes of atmospheric circulation variability when the climate system is subjected to time-dependent external forcing, via introducing the snapshot empirical orthogonal function (SEOF) analysis. Rather than apply empirical orthogonal function (EOF) analysis in the traditional temporal dimension we compute instantaneous EOFs (spatial patterns of variability) and associated principal components (PCs, amplitude and polarity of the patterns) across the ensemble dimension. In doing so, we can monitor the changes in an EOF mode resulting from the time-dependent external forcing and account for the non-stationarity of internal variability. We note that a similar method was also developed in Maher et al. (2018), however, that approach combines the variability of the monthly data with that of the ensemble.

The instantaneous strength of the linkage between a particular SEOF teleconnection pattern and another climate variable (e.g., surface temperature, TS) can be quantified by means of the *snapshot correlation*, i.e., the instantaneous Pearson correlation coefficient computed across the ensemble. In this way, instantaneous correlation maps are obtained, thereby allowing one to monitor the spatial distribution of the correlation field in tandem with its time evolution. Such an approach has been insightful for documenting changes in the teleconnections of the ENSO (Bódai et al. 2020, Haszpra et al. 2020a) and that of the North Atlantic Oscillation (Herein et al. 2017). Similar to EOF analysis, maximum covariance analysis (MCA, Bretherton et al. 1992) may also analogously be extended to its ensemble-based twin (snapshot MCA, SMCA) to study forced changes in coupled modes of variability (see below).

Results

An illustrative example of SEOF: The Arctic Oscillation

We briefly demonstrate advantages of SEOF analysis in monitoring temporal changes in the Arctic Oscillation (AO) under RCP8.5 forcing in the CESM Large Ensemble (CESM-LE, Kay et al. 2015) for 1950–2099. We define the AO as the leading SEOF mode in the winter (December–January–February, DJF) seasonal mean sea level pressure (SLP) anomalies poleward of 20°N and the corresponding PC series as the instantaneous (DJF) AO indices (AOI). Thus, for each winter season during 1950–2099 we obtain a spatial pattern that characterizes the current set of potential climate states (spanned by the ensemble spread) and explains the largest variability in their SLP fields, in addition to a PC series whose length is the number of ensemble members (for the AO, this PC is termed the AO Index or AOI).

The left panel in Figure 1 illustrates the instantaneous DJF mean SLP anomalies regressed onto the leading SEOF mode in 2025 in the CESM-LE, which closely resembles the observed AO pattern (based on historical reanalysis, not shown but see for example Thompson et al. 2000). Repeating the SEOF analysis for each year between 1950–2099, important characteristics of the model's AO are revealed, including temporal changes in amplitude at several locations determined from a linear fit to the regression values at each grid box (Figure 1 right). For clarity, in this panel dots represent geographical locations

where the linear trend is found to be significant at the 95% level and crosses indicate where, in addition, the regression coefficients are significant at the 95% level in the temporal mean. For example, the amplitude of the AO in the Pacific center-of-action shows an increase of about 0.02 hPa yr^{-1}, implying 3 hPa over 150 years, i.e., the change is of the same magnitude as the typical amplitude in 2025 (2.5–6.5 hPa). On the contrary, in the northern part of Europe and Asia a slight decrease of the AO amplitude can be observed. Moreover, the choice of the scenario influences the magnitude of the changes (see application to the MPI Grand Ensemble (MPI-GE) with three forcing scenarios in Haszpra et al. 2020b).

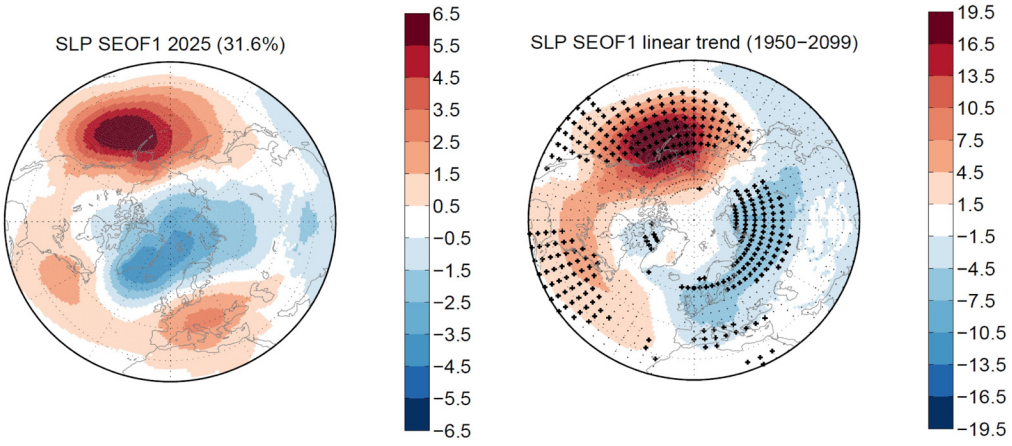


Figure 1. (left) December-January-February (DJF) mean sea level pressure (SLP) anomalies (hPa) regressed onto the first SEOF mode (explained variance is indicated in parenthesis) in 2025 based on the CESM-LE (RCP8.5 scenario). (right) Linear trend (10-3 hPa yr-1) in the SEOF SLP regression coefficients during 1950-2099 based on the CESM-LE under historical and RCP8.5 forcing. Dots represent geographical locations where the trend is significant at the 95% level. Crosses indicate where, in addition, the regression coefficients are significant at the 95% level in the temporal mean.

The AOIs are constructed for each winter season during 1950–2099 by projecting the instantaneous SLP anomalies of the ensemble members onto the given (instantaneous) loading (SEOF) pattern. A rather straightforward step is to calculate the snapshot correlation coefficient *r* field between the AOI and the surface air temperature (T_S) across the ensemble. In the left panel of Figure 2, the correlation map is shown for 2025. Similar to the SLP regression map, the correlation map resembles the observed relationship between the AO and T_S (Wallace and Gutzler 1981). Fitting a linear trend at each grid point to the time series of snapshot correlation coefficients, significant changes are evident in the strength of the teleconnections across the NH. Dots and crosses in the right panel of Figure 2 indicate regions where the snapshot correlation coefficients undergo significant changes over time, and may need to be taken into account in future seasonal prediction. These regions include, e.g., Alaska, the eastern part of the Pacific Ocean and Northern Europe where the negative correlations become more pronounced (the correlation coefficient decreases by 0.1–0.3 over 150 years), and a substantial positive trend can be found in the eastern part of Asia where the correlation coefficient increases from about 0.6 to 0.8 over 150 years.

An illustrative example for SMCA: Atmosphere–sea ice coupling under different forcing scenarios

Next, we consider an example from a different season and study how the coupling between the summertime (June–July–August, JJA) Arctic atmospheric circulation and September sea ice variability might depend on future anthropogenic forcing (Ding et al. 2019). Concomitant patterns of high-pressure anomalies in the Arctic and enhanced sea ice melt has been previously shown in SMILEs (Topál et al. 2020), but there has been little focus on possible nonlinearities in the nature or strength of the coupling.

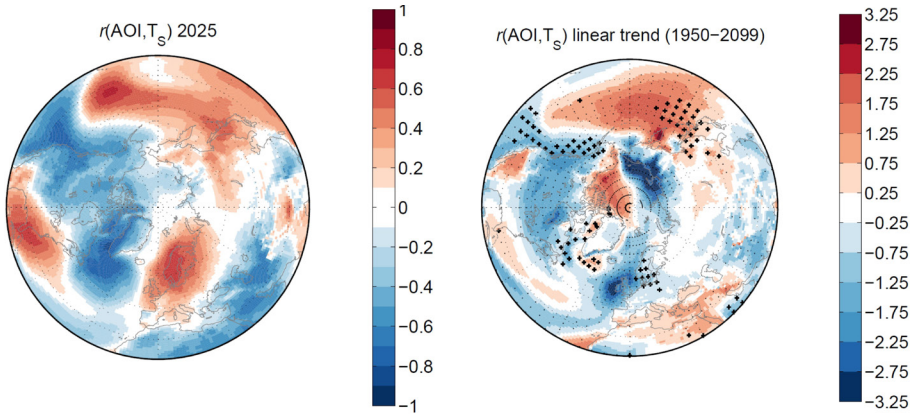


Figure 2. (left) Ensemble-based snapshot correlation coefficient field between the Arctic Oscillation Index and surface air temperature in 2025 based on the CESM-LE (RCP8.5 scenario). (right) Linear trend (10^{-3} yr^{-1}) in the snapshot correlation coefficient during 1950–2099 based on the CESM-LE under historical an RCP8.5 forcing. Dots represent geographical locations where the trend is significant at the 95% level. Crosses indicate where, in addition, the correlation coefficients are significant at the 95% level in the temporal mean.

To study the influence of external forcing on the coupling between the atmospheric circulation and sea ice, we calculate linear trends in all members of each of the three RCP scenarios in MPI-GE (Maher et al. 2019) over 2020–2050 for both JJA 200-hPa geopotential height (Z200) and September sea ice concentration (SIC) within the Arctic (poleward of 60°N). Second, we remove the ensemble mean trend from each member, so the residual trends only reflect internal variability of the model over the 31 years. We then use SMCA between JJA Z200 and September SIC trend fields across all the members in a given scenario. The leading ensemble-based SMCA modes reflect the dominant coupled patterns of internally-generated trends in Z200 and SIC. For comparison, we also calculate the September sea ice area (SIA) trends in each member. We note, that a similar approach, using ensemble member trend-based EOFs, has previously been presented in Wettstein and Deser (2014) to study co-variability of atmospheric circulation and sea ice.

The extent to which the SIA trends resemble the time expansion coefficients of SMCA in each member can be used to probe the degree of linearity in the coupling between Z200 and SIC in a given RCP scenario. In the case of the RCP4.5 scenario, nearly half of the members in the fast melting group (15% of the members with strongest sea ice melt) show identically strong sea ice melt despite the linear decrease in the time expansion coefficient series of the same members (Figure 3d). Such a phenomenon is not observed under the RCP2.6 and 8.5 scenarios (Figure 3a,g), which suggests that the coupling may exhibit stronger non-linearity under the RCP4.5 scenario. We also show that the spatial patterns of Z200 and SIC corresponding to the leading ensemble-based SMCA mode differs slightly between RCP4.5 and the two other forcing scenarios, indicating some role for the intensity of external forcing, which remains a subject of future work (Figure 3b-c, e-f, h-i). Interestingly, the shared fraction of co-variance between Z200 and SIC (indicated in the panel titles in Figure 3) are also slightly higher for the RCP4.5 scenario compared to the other two. Regarding the physical mechanism behind the observed co-variability between atmospheric circulation and sea ice, we argue based on previous work (Ding et al. 2017; Baxter et al. 2019) that an internal atmospheric process manifested as a high-pressure driven adiabatic warming (via regulating downward longwave radiation) can cause sea ice melt on top of the externally forced melting (Figures 3b-c, e-f, h-i). A more thorough discussion of this physical mechanism and its limited representation in SMILEs can be found in Topál et al. (2020).

Outlook

We have applied the mathematically well-established "snapshot view" based on dynamical systems theory to the analysis of SMILEs and reconsidered traditional methodologies to study possible future changes in internal variability. A future direction of the research

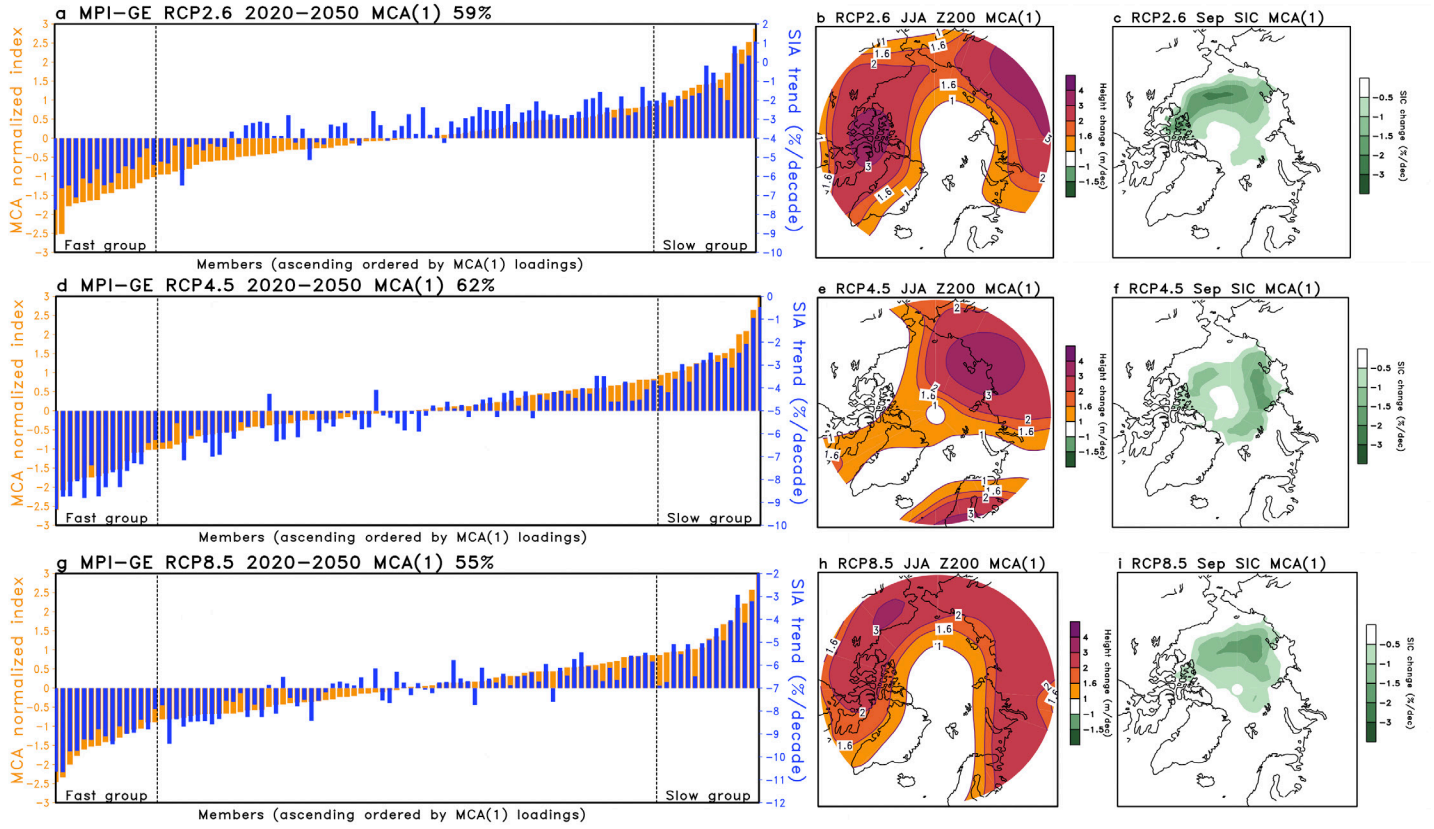


Figure 3. Snapshot maximum covariance analysis (SMCA) between June-July-August Z200 and September sea ice concentration (SIC) trends during 2020–2050 in the three RCP scenarios (2.6, 4.5 and 8.5) from the MPI-GE. The bar plots on the left represent the normalized member loadings (orange bars: left y axis) of the leading SMCA mode and the corresponding members' September total sea ice area (SIA) trends (blue bars: right y axis), arranged in ascending order of the SMCA loadings. The percentages represent the shared fraction of covariance between Z200 and SIC trends across each ensemble explained by the leading mode. Maps in the middle and right columns show the Z200 and SIC spatial patterns, respectively, of the leading SMCA mode for each scenario.

could be comparing SEOF results to observations. One might try a comparison using carefully chosen (but still subjective) multiple time windows centered to the instantaneous SEOF year to construct the relevant traditional EOF pattern using a single time series. However, this comparison is expected to yield similar results only if the external forcing does not change much within the chosen time window and, therefore, ergodicity approximately holds. Equally, it is to be noted that the ongoing climate change is not ergodic (Tél et al. 2019). As a consequence, the above-mentioned comparison can serve as a measure of the ergodicity as well. A crucial message of the snapshot view is that all of the traditional, time series-based methods can be reformulated for ensembles, which will be of use for the broader climate community. In this way, utilizing ensemble-based (snapshot) analyses of the available SMILEs, ambiguous results arising from subjective choices of traditional methods (e.g. length and center of time windows) can be avoided, the possible climate states at each time instant can be properly characterized, and forced changes in any ensemble-based quantity can be determined.

Acknowledgments

This article was supported by the ÚNKP-18-4 (T.H.) and ÚNKP-19-3 (D.T.) New National Excellence Program of the Ministry for Innovation and Technology, by grant NTP-NFTÖ-18 (D.T.) of the Ministry of Human Capacities, by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (T.H., M.H.), and by the National Research, Development and Innovation Office–NKFIH under Grants PD-121305 and PD-132709 (T.H.), PD-124272 (M.H.), FK-124256 and K-125171 (T.H., M.H.). The authors are thankful to the developer group of MPI-GE for providing the MPI-GE ensembles and wish to thank the Climate Data Gateway at NCAR for providing access to the output of the CESM-LE. The authors acknowledge and especially thank Gábor Drótos, Tamás Bódai and János Márfy for establishing the fundamentals of applying snapshot attractors theory to climate dynamics, including early stages of their SEOF-related research. The authors are grateful to Tamás Tél for inspiring ideas and suggestions on the subject of the section Theoretical background.

References

Barnes, E. A. and L. Polvani, 2013: Response of the midlatitude jets, and of their variability, to increased greenhouse gases in the CMIP5 models. *J. Climate*, **26**, 7117–7135, doi:10.1175/JCLI-D-12-00536.1.

Baxter, I., and Coauthors, 2019: How tropical Pacific surface cooling contributed to accelerated sea ice melt from 2007 to 2012 as ice is thinned by anthropogenic forcing. *J. Climate*, **32**, 8583–8602, doi:10.1175/JCLI-D-18-0783.1.

Bódai T., and T. Tél, 2002: Annual variability in a conceptual climate model: Snapshot attractors, hysteresis in extreme events, and climate sensitivity. *Chaos*, **22**, 023110, doi:10.1063/1.3697984.

Bódai, T., G. Drótos, M. Herein, F. Lunkeit, and V. Lucarini, 2020: The forced response of the El Niño–Southern Oscillation–Indian Monsoon teleconnection in ensembles of Earth system models. *J. Climate*, **33**, 2163–2182, doi:10.1175/JCLI-D-19-0341.1.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, **5**, 541–560, doi:10.1175/1520-0442(1992)005<0541:AIOMFF>2.0.CO;2.

Chekroun M. D., E. Simonnet, and M. Ghil, 2011: Stochastic climate dynamics: Random attractors and time dependent invariant measures. *Phys. D*, **240,** 1685, doi:10.1016/j.physd.2011.06.005.

Deser, C., A. Phillips, V. Bourdette, H. Teng, 2012: Uncertainty in climate change projections: the role of internal variability. *Clim. Dyn.* **38,** 527–546, doi:10.1007/s00382-010-0977-x.

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Clim. Change*, **10**, 277–286, doi:10.1038/s41558-020-0731-2.

Ding, Q., and Coauthors, 2019: Fingerprints of internal drivers of Arctic sea ice loss in observations and model simulations. *Nat. Geosci.,* **12**, 28–33, doi:10.1038/s41561-018-0256-8.

Ding, Q., and Coauthors, 2017: Influence of high-latitude atmospheric circulation changes on summertime Arctic sea ice. *Nat. Climate Change*, **7**, 289–295, doi:10.1038/nclimate3241.

Drótos, G., T. Bódai, and T. Tél, 2015: Probabilistic concepts in a changing climate: snapshot attractor picture. *J. Climate*, **28**, 3275–3288, doi:10.1175/JCLI-D-14-00459.1.

Ghil M, M. D. Chekroun, and E. Simonnet, 2008: Climate dynamics and fluid mechanics: Natural variability and related uncertainties. *Phys. D*, **237**, 2111, doi:10.1016/j.physd.2008.03.036.

Haszpra, T., M. Herein, and T. Bódai, 2020a: Investigating ENSO and its teleconnections under climate change in an ensemble view – a new perspective. *Earth Syst. Dyn*., **11**, 267–280, doi:10.5194/esd-11-267-2020.

Haszpra, T., D. Topál, and M. Herein, 2020b: On the time evolution of the Arctic Oscillation and related wintertime phenomena under different forcing scenarios in an ensemble approach. *J. Climate*, **33**, 3107–3124, doi:10.1175/JCLI-D-19-0004.1.

Hawkins, E., and R. Sutton, 2009: The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bull. Amer. Meteor. Soc*., **90**, 1095–1108, doi:10.1175/2009BAMS2607.1.

Hawkins, E., D. Frame, L. Harrington, M. Joshi, A. King, M. Rojas, R. Sutton, 2020: Observed emergence of the climate change signal: from the familiar to the unknown. *Geophys. Res. Lett*., **47**, e2019GL086259, doi:10.1029/2019GL086259

Herein M., G. Drótos, T. Haszpra, J. Márfy, and T. Tél, 2017: The theory of parallel climate realizations as a new framework for teleconnection analysis. *Sci. Rep*. **7**, 44529, doi:10.1038/srep44529.

Herein, M., J. Márfy, G. Drótos, and T. Tél, 2016: Probabilistic concepts in intermediate-complexity climate models: A snapshot attractor picture. *J. Climate*, **29**, 259–272, doi:10.1175/JCLI-D-15-0353.1.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) Large Ensemble Project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc*., **96**, 1333–1349, doi:10.1175/BAMS-D-13-00255.1.

Ku, W. L., M. Girvan, and E. Ott, 2015: Dynamical transitions in large systems of mean field-coupled Landau-Stuart oscillators: Extensive chaos and cluster states. *Chaos,* **25**, 123122, doi:10.1063/1.4938534.

Lehner, F., C. Deser, N. Maher, J. Marotzke, E. M. Fischer, L. Brunner, R. Knutti, E. Hawkins, 2020: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6, *Earth Syst. Dyn*., **11**, 491–508, doi:10.5194/esd-11-491-2020

Leith, C, 1978: Predictability of climate. *Nature*, **276**, 352–355, doi:10.1038/276352a0.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Maher, N., and Coauthors, 2019: The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst*., **11**, 2050– 2069, doi:10.1029/2019MS001639.

Maher, N., D. Matei, S. Milinski, and J. Marotzke, 2018: ENSO change in climate projections: Forced response or internal variability? *Geophys. Res. Lett*., **45**, 11,390-11,398, doi:10.1029/2018GL079764.

Merrifield, A. L., and Coauthors 2019: A weighting scheme to incorporate large ensembles in multi-model ensemble projections, *Earth Syst. Dyn. Discuss*. doi:10.5194/esd-2019-69

Ott, E, 1993: *Chaos in dynamical systems*. Cambridge University Press. 478 pp.

Peters, G. P., R. M. Andrew, J. G. Canadell, P. Friedlingstein, R. B. Jackson, J. I. Korsbakken, C. Le Quéré, A. Peregon, 2020: Carbon dioxide emissions continue to grow amidst slowly emerging climate policies, *Nat. Clim. Change,* **10**, 3–6, doi:10.1038/s41558-019-0659-6

Pierini, S., 2014: Ensemble simulations and pullback attractors of a periodically forced double-gyre system. *J. Phys. Oceanogr*., **44**, 3245–3254, doi:10.1175/JPO-D-14-0117.1.

Pikovsky, A. S., 1984: Synchronization and stochastization of the ensemble of autogenerators by external noise. *Radiophys. Quantum Electron*., **27**.

Romeiras F. J., C. Grebogi, and E. Ott, 1990: Multifractal properties of snapshot attractors of random maps, *Phys. Rev. A,* **41**, 784, doi:10.1103/PhysRevA.41.784.

Screen, J, 2014: Arctic amplification decreases temperature variance in northern mid- to high-latitudes. *Nat. Clim. Change,* **4,** 577–582, doi:10.1038/nclimate2268

Screen, J., and Simmonds, I, 2010: The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature,* **464,** 1334–1337, doi:10.1038/nature09051

Serquina, R., Y. C. Lai, and Q. Chen, 2008: Characterization of nonstationary chaotic systems, *Phys. Rev. E,* **77**, 26208, doi:10.1103/PhysRevE.77.026208.

Tél, T. and Coauthors, 2019: The theory of parallel climate realizations: A new framework of ensemble methods in a changing climate – an overview. *J. Stat. Phys*., doi:10.1007/s10955-019-02445-7.

Tél, T., and M. Gruiz, 2006: *Chaotic dynamics*. Cambridge University Press. 393 pp.

Topál, D., Q. Ding, J. Mitchell, I. Baxter, M. Herein, T. Haszpra, R. Luo, and Q. Li, 2020: An internal atmospheric process determining summertime Arctic sea ice melting in the next three decades: Lessons learned from five large ensembles and multiple CMIP5 climate simulations. *J. Climate*, **33**, 7431–7454, doi:10.1175/JCLI-D-19-0803.1.

Vincze M., I. D. Borcia, and U. Harlander, 2017: Temperature fluctuations in a changing climate: an ensemble based experimental approach. *Sci. Rep*. **7**, 254, doi:10.1038/s41598-017-00319-0.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev*., **109**, 784–812, doi:10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2.

Wettstein, J. J., and C. Deser, 2014: Internal variability in projections of twenty-first-century Arctic sea ice loss: role of the large-scale atmospheric circulation. *J. Climate*, **27**, 527–550, doi:10.1175/JCLI-D-12-00839.1.

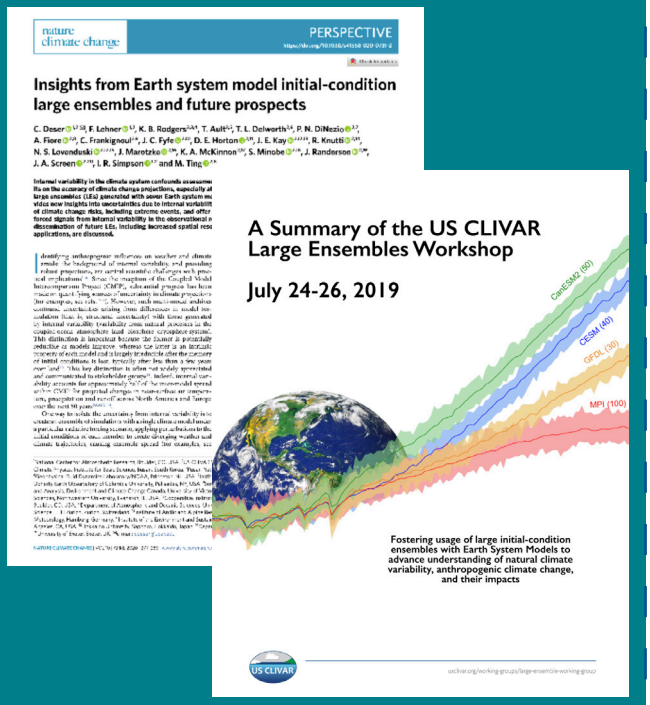
Yu L., E. Ott, and Q. Chen, 1990: Transition to chaos for random dynamical systems, *Phys. Rev. Lett.*, **65**, 2935, doi:10.1103/PhysRevLett.65.2935.

Want more information on using Large Ensembles?

Insights from Earth system model initial-condition large ensembles and future prospects (Deser et al. 2020)

2019 Large Ensembles Workshop Summary (Large Ensemble Working Group 2020)

Coming in September: NCAR Climate Variability Diagnostics Package for Large Ensembles



ANNOUNCEMENTS

Upcoming US CLIVAR opportunities

2020 Fall Workshop Call

Coming in August: The US CLIVAR program sponsors open community workshops, conferences, and science meetings to coordinate, develop, plan, and implement new or focused activities for the benefit of the scientific community and relevant to the goals of US CLIVAR. Workshops can serve as an initiation point in the planning process for future community activities. Funding is limited and not all submitted workshop requests may be supported.

2020 Working Group Call

Coming in September: The US CLIVAR program establishes limited-lifetime, action-oriented Working Groups of scientists to coordinate and implement focused activities for the benefit of the scientific community. Working Groups supported by this year's call will be initiated by Spring 2021 and undertake actionable and measurable tasks over a 2–3 year period.

2021 Panel Member Nominations

Coming in November: The US CLIVAR Scientific Steering Committee seeks qualified individuals to serve on its three Panels to enhance current strengths while adding expertise in new areas. Qualified nominees are expected to represent the broader interests of the research community, be willing and able to engage in scientific as well as programmatic discussions to advance Panel activities, and work with other members of the US and international CLIVAR communities.

Stay up to date with all the latest US CLIVAR news by subscribing to the mailing list

usclivar.org
uscpo@usclivar.org
twitter.com/usclivar

US Climate Variability and Predictability (CLIVAR) Program
Washington, DC

US CLIVAR acknowledges support from these US agencies:

This material was developed with federal support of NASA (80NSSC17M0007), NOAA (NA11OAR4310213), NSF (AGS-1502208), and DOE (DE-SC0016332). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agencies.