VARIATIONS

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.

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.
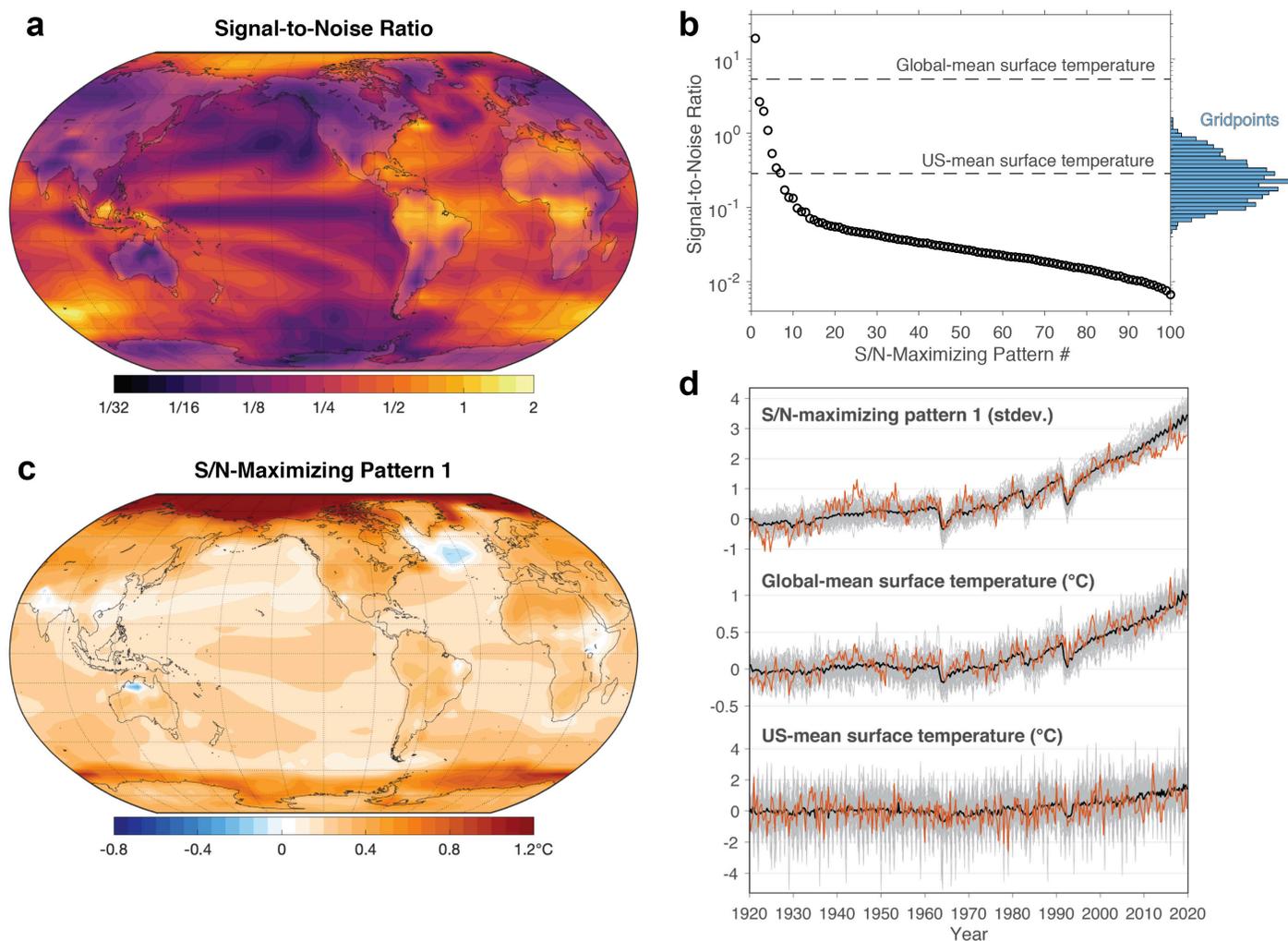
Figure 1. (a) Signal-to-noise ratio of seasonal (3-month-average) surface temperature anomalies in the CESM Large Ensemble (Kay et al. 2015) over the period 1920-2019. (b) Signal-to-noise ratio of signal-to-noise-maximizing (S/N-maximizing) patterns (Ting et al. 2009; Wills et al. 2020) of CESM-LE and 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
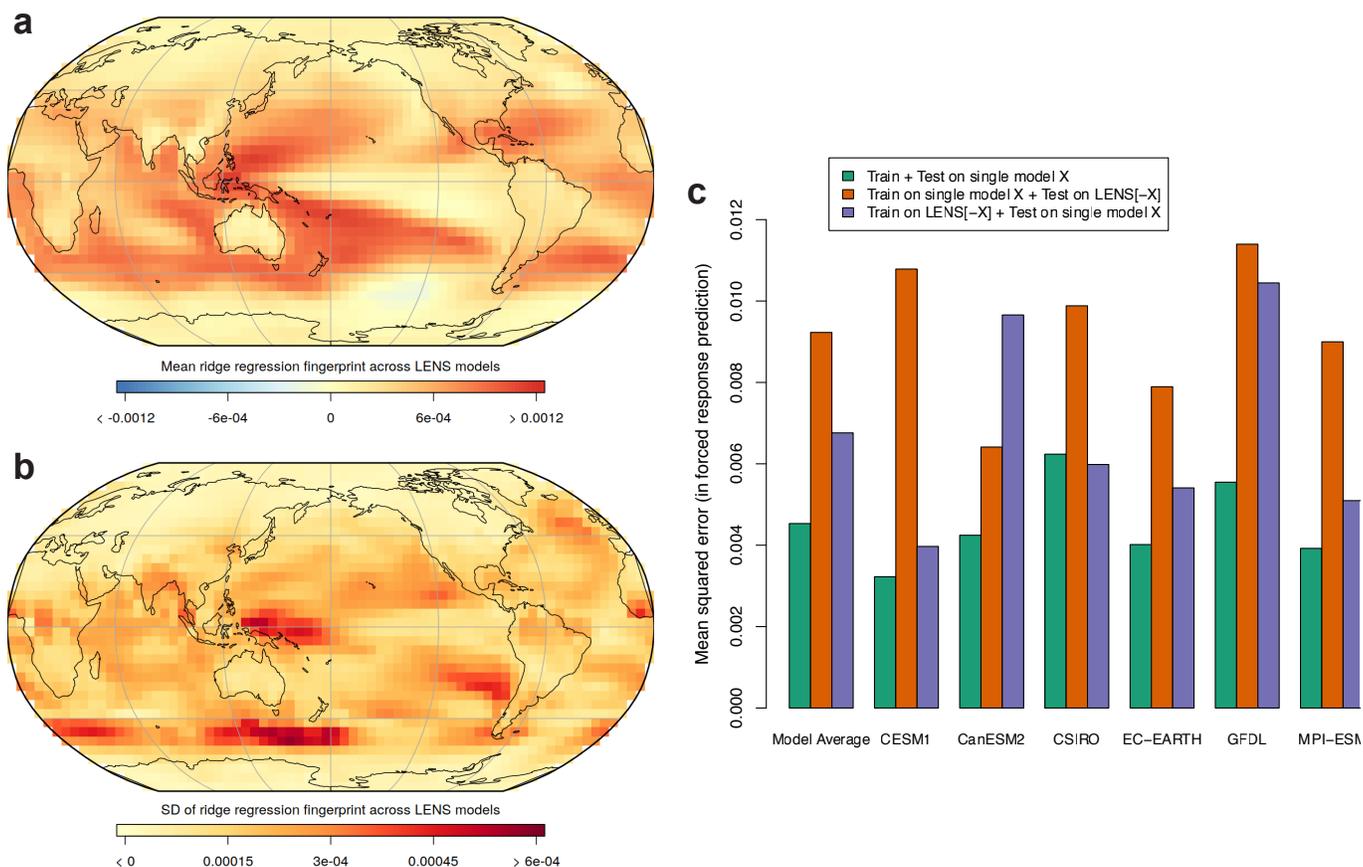
Figure 2. Illustration of the fingerprint for predicting the forced global-mean surface temperature response in the form of regression coefficients averaged over six different SMILEs for a fixed λ value (log10(λ) = 1.06). (b) Standard deviation of regression coefficients across different SMILEs, highlighting 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 $\log10(\lambda) = 1.06$, 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.

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<1957:OFFTDO>2.0.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.