

Predicting when and where climate change becomes robustly detectable

Ramdane Alkama

ram.alkama@hotmail.fr

Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE)

Jerome Ogee

INRAE, Unité Mixte de Recherche (UMR)1391 ISPA

Didier Swingedouw

EPOC, Universite Bordeaux 1

Giovanni Forzieri

University of Florence, Dipartimento di Ingegneria Civile e Ambientale

Herve Douville

Météo-France

Bertrand Dechame

CNRM-Meteo France

Christine Delire

CNRM, Université de Toulouse, Météo-France, CNRS

Fabio D'Andrea

Laboratoire de Météorologie Dynamique

Gilles Ramstein

LSCE

Ayoub Zeroual

Ecole Nationale Supérieure de l'Hydraulique

Article

Keywords:

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11116199/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Predicting when and where climate change becomes robustly detectable

Ramdane Alkama^{1*}, Jérôme Ogée¹, Didier Swingedouw², Giovanni Forzieri³, Hervé Douville⁴, Bertrand Decharme⁴, Chrestine Delire⁴, Fabio D'Andrea⁵, Gilles Ramstein⁶ & Ayoub Zeroual⁷

¹ Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), Unité Mixte de Recherche (UMR)1391 ISPA, Villenave D'Ornon, France

² Environnements et Paléoenvironnements Océaniques et Continentaux (EPOC), Université de Bordeaux, Pessac, France

³ Department of Civil and Environmental Engineering, University of Florence, Florence 50139, Italy

⁴ Météo-France, CNRS, Univ. Toulouse, CNRM, Toulouse, France

⁵ Laboratoire de Météorologie Dynamique, IPSL, CNRS, Paris, France

⁶ Laboratoire des Sciences Du Climat et de L'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ Université Paris-Saclay, Gif-sur-Yvette, France

⁷ Water Engineering and Environment Laboratory, National Higher School of Hydraulics (ENSH-Blida), Blida 09000, Algeria

* Corresponding author. Email: ram.alkama@hotmail.fr

Abstract

Regional climate responses can remain difficult to identify because internal variability can mask externally forced changes for decades. Here we derive an analytical scaling law showing that recoverability, the expected agreement between a sampled climate response and its underlying forced component, is determined by only two quantities: the regional signal-to-noise ratio and the number of independent realizations. Across large climate-model ensembles, the relationship predicts both the agreement attainable for a given ensemble size and the sampling required to reach a prescribed agreement, without empirical calibration. Observations of temperature, precipitation and sea-level pressure follow the same relationship in the single-realization limit, linking ensemble convergence and model-observation agreement within a common framework. Applied through time, the scaling predicts where and when regional forced responses become recoverable as signal-to-noise ratios evolve, providing a quantitative basis for assessing the robustness of regional climate information.

37 Main

38 Anthropogenic climate change is now detectable throughout the climate system, enabling
39 increasingly rigorous comparisons between simulated and observed regional climate
40 responses^{1,2}. Yet regional climate change remains considerably more difficult to interpret than
41 global warming because externally forced responses are often obscured by internally generated
42 variability over several decades^{3–6}. Consequently, regional changes in precipitation,
43 atmospheric circulation and other hydroclimatic variables frequently appear inconsistent across
44 climate models, individual realizations and observations^{7–12}. Distinguishing genuine
45 differences in externally forced climate responses from uncertainty arising through finite
46 sampling is however essential for climate model evaluation, impact assessment and climate
47 adaptation.

48 Several complementary statistical frameworks have been developed to address different aspects
49 of this problem. Large ensembles ask how many realizations are required to recover externally
50 forced climate responses from internal variability^{3,13–16}. Model evaluation asks whether
51 apparent disagreements between simulations and observations arise from structural model
52 deficiencies or simply from the limited sampling of a noisy climate system^{7,17–19}. Climate
53 emergence asks when externally-forced climate change becomes distinguishable from internal
54 variability in the single realization provided by the real climate system^{6,20–23}. Meanwhile,
55 signal-to-noise ratio (SNR) has long been recognized as a fundamental measure of climate
56 predictability, emergence and forecast skill^{24–28}.

57 Here, we explore whether these three approaches, that all seek to recover externally-forced
58 regional climate responses from finite and noisy climate information, could represent different
59 manifestations of the same statistical relationship. If so, seemingly distinct questions, including
60 how many realizations are required to recover a forced response, how much agreement should
61 be expected between models and observations, and when regional climate change becomes
62 robustly identifiable, may ultimately have a common answer. Such a framework would provide
63 a unified basis for interpreting regional climate uncertainty across climate models, observations
64 and future projections.

66 Can externally forced regional climate responses be recovered?

67 Large initial-condition ensembles have become the standard approach for estimating forced
68 regional climate responses because averaging across realizations progressively suppresses
69 internal variability^{3–5,13,29}. Consistent with this view, individual climate-model realizations
70 recover externally forced regional precipitation responses only incompletely (Fig. 1a).
71 Agreement with the independently-estimated forced response remains weak over large parts of
72 the globe, particularly across western boundary current systems, tropical transition regions and
73 mid-latitude storm tracks, where internal variability dominates regional precipitation changes.

74 Ensemble averaging substantially improves the recovery of the forced response. Despite
75 differences in model formulation, ensemble size and forcing implementation, the ten largest
76 CMIP6 Single-Model Initial-condition Large Ensembles (SMILEs) recover remarkably similar
77 large-scale precipitation responses, including high-latitude wetting and subtropical drying (Fig.
78 1b). This consistency extends to each individual SMILE (Extended Data Fig. 1), indicating that
79 the recovered signal is not controlled by any single model but represents a common externally

forced response. Models with larger ensembles recover this response more completely. The large-scale organization is also robust to trend estimation and detrending (Extended Data Figs. 2-4).

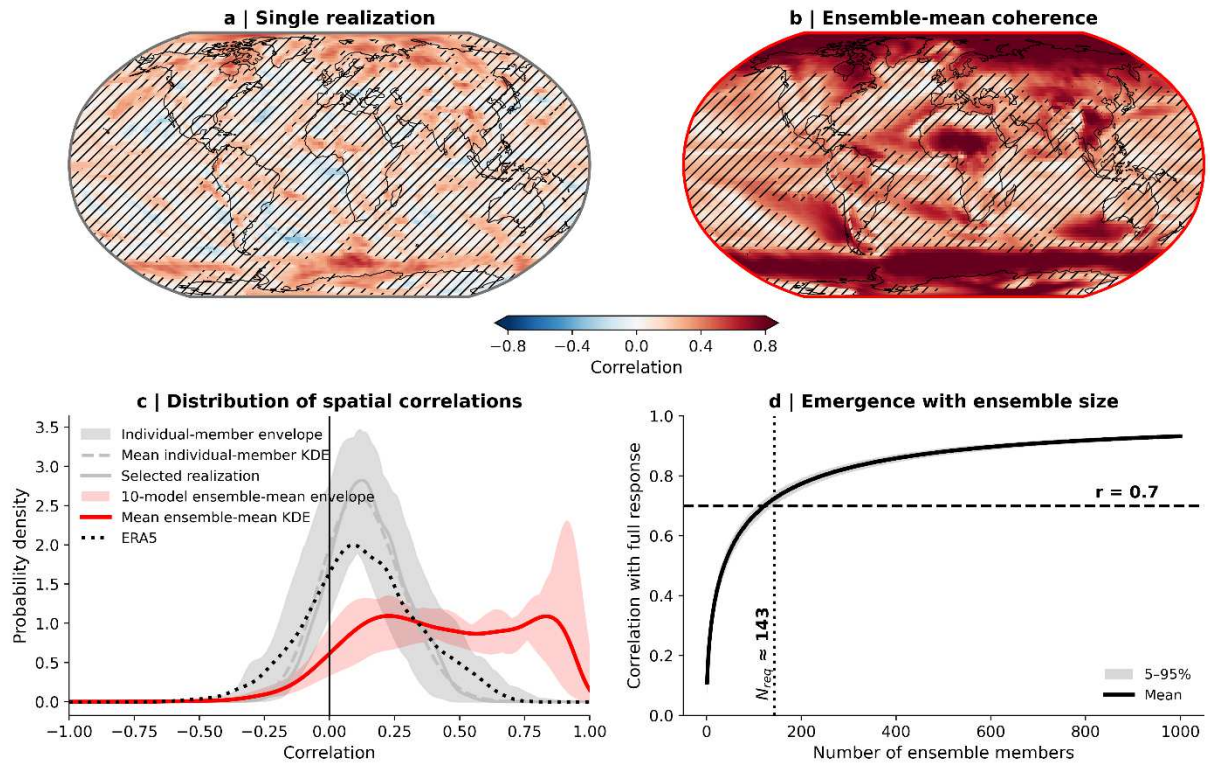


Figure 1 Recovering coherent regional forced climate responses through ensemble averaging. a) Temporal correlations between a single historical precipitation realization and the independently estimated multi-model forced response over 1940-2014. Hatching denotes correlations that are not statistically significant ($P > 0.05$). b) Temporal correlations between the ensemble-mean response of the ten largest CMIP6 single model initial-condition large ensembles (including CESM2-LE) and the corresponding leave-one-model-out multi-model forced response. For each model, the reference response is estimated from all remaining models to ensure statistical independence. Ensemble means are weighted according to the number of realizations contributed by each model. Hatching indicates regions where fewer than 90% of the member-weighted model contributions exhibit positive and statistically-significant correlations ($P < 0.05$). c) Probability density distributions of temporal correlations. Grey shading denotes the range across all individual realizations, the dashed grey curve their mean distribution, and the solid grey curve the realization shown in a. Red shading denotes the range across the ten leave-one-model-out ensemble means, with the red curve indicating their member-weighted mean distribution. The black dotted curve shows the corresponding distribution for ERA5 observations. d) Bootstrap estimates of the recovery of the forced precipitation response as ensemble size increases. The black curve shows the mean correlation with the independently estimated forced response, grey shading the 5th-95th percentile range. The vertical dotted line indicates the ensemble size required to reach a prescribed level of correlation agreement of $r = 0.7$.

The contrast between individual realizations and ensemble means is also reflected in the distribution of regional agreement (Fig. 1c). Individual climate-model realizations span a broad range of generally weak correlations with the forced response, whereas ensemble means cluster around substantially higher correlation values. Importantly, the observed agreement derived

from ERA5 lies well within the distribution of individual climate-model realizations rather than ensemble means, indicating that observations exhibit the level of agreement expected from a single realization of the climate system. This result is consistent with previous studies showing that a single observational record cannot be expected to reproduce the forced regional response exactly because it contains an irreducible contribution from internally generated variability^{4,6,30}.

Bootstrap resampling confirms that agreement with the forced response increases monotonically as additional realizations are averaged (Fig. 1d), providing an empirical estimate of the sampling required to recover the forced response. Together, these results motivate a quantitative relationship linking recoverability to local SNR and sampling.

Can regional climate recoverability be predicted?

To explain the empirical increase in recoverability with ensemble size (Fig. 1d), we derive a simple analytical scaling law linking the agreement with the underlying forced response (r) to only two quantities: the local SNR and the number of independent realizations or sampling size (N) (see Methods):

$$r = \left(1 + \frac{1}{N(SNR)}\right)^{-1/2} \quad (\text{Eq 1})$$

For any combination of SNR and sampling size, Eq. 1 predicts the expected agreement with the forced response; conversely, prescribing a target agreement determines the required sampling (Fig. 2a). We evaluate this prediction against bootstrap estimates of the sampling required to robustly reach $r=0.7$ (see Methods and Fig. 1d). Required sampling size (N_{req}) varies by more than an order of magnitude across regions (Fig. 2b). Some regions recover the forced response from only a few realizations, whereas circulation-sensitive regions require substantially larger ensembles to achieve the same level of agreement. The geographical distribution of N_{req} closely mirrors regional variations in $1/SNR$ (Fig. 2c), in agreement with Eq. 1.

The analytical scaling law accurately reproduces these empirical results. The predicted relationship between N_{req} and the inverse SNR closely follows the bootstrap estimates (Fig. 2d), and the analytical predictions $N_{th} = \frac{r^2}{1-r^2} \frac{1}{SNR}$ agree closely with independently estimated bootstrap values across the full range of sampling requirements (Fig. 2e). Although N_{req} is estimated entirely empirically from bootstrap resampling, its geographical distribution and magnitude are accurately predicted by a simple analytical expression depending only on SNR. Rather than simply confirming that larger ensembles improve estimates of externally forced climate change, a result well established by previous large-ensemble studies^{3,13,29}, the scaling law explains why sampling requirements vary geographically and quantitatively predicts the sampling requirement from the local balance between forced signal and internally generated variability. The generality of the analytical relationship is confirmed across multiple independent tests. The same scaling law is recovered from bootstrap experiments, leave-one-model-out analyses and for three contrasting climate variables (precipitation, surface temperature and seal-level pressure, see Extended Data Figs. 5-9), demonstrating that it is robust to the estimation method, the climate variable and the ensemble used to derive the forced response.

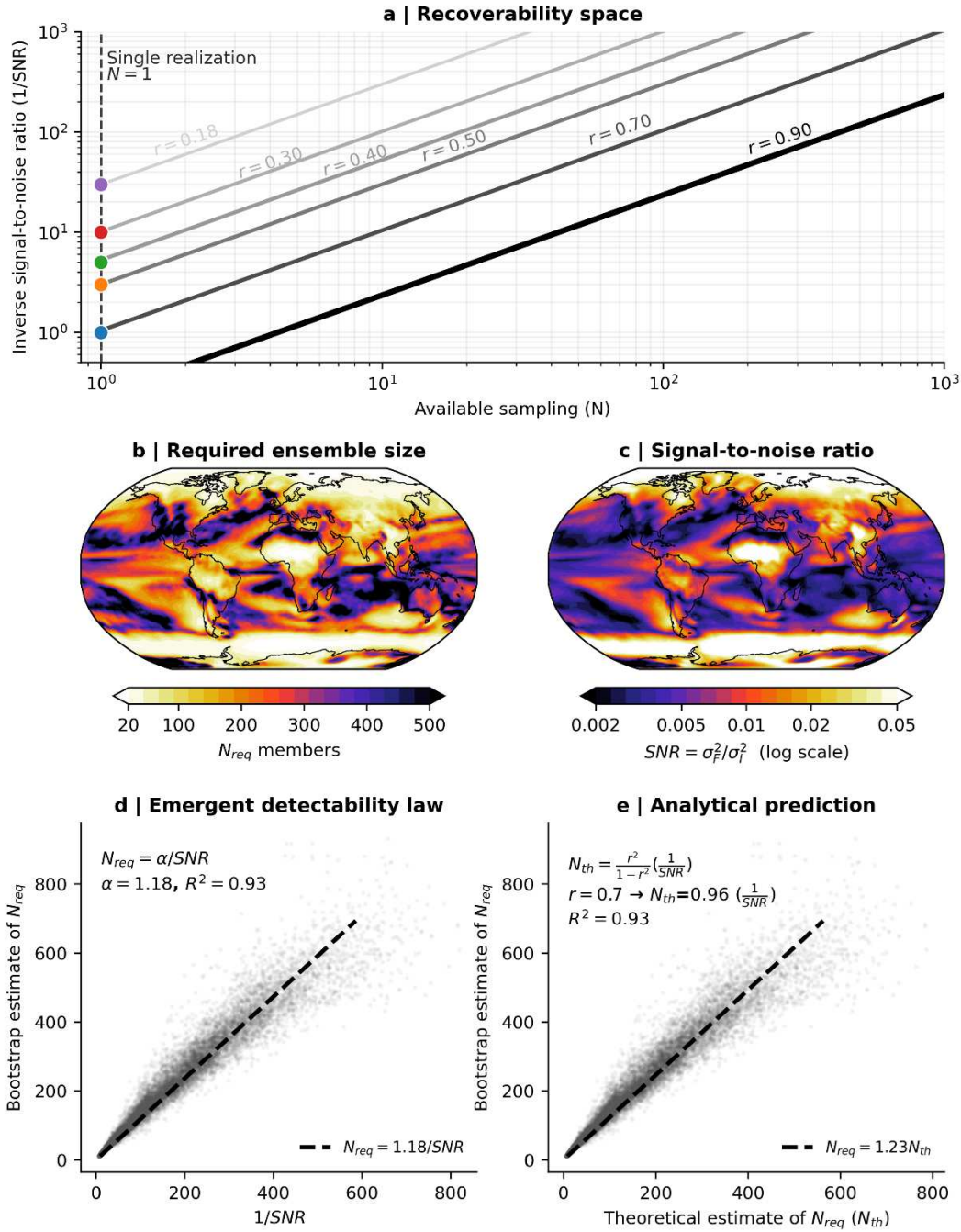


Figure 2 Regional climate recoverability is governed by a simple analytical relationship. a) Conceptual representation of the theoretical recoverability framework: $r = (1 + \frac{1}{N(SNR)})^{-1/2}$ linking attainable agreement with the externally forced response (r), available sampling (N), and the local signal-to-noise ratio (SNR). b) Spatial distribution of the required ensemble size (N_{req}) to recover the forced precipitation response with a target agreement of $r = 0.7$ over 1940-2014. At each grid point, (N_{req}) is estimated from 1,000 bootstrap resampling experiments with replacement as the smallest ensemble size for which the 5th percentile of the bootstrap correlation distribution is greater than or equal to 0.7, ensuring that at least 95% of bootstrap estimates satisfy the prescribed agreement criterion. c) Spatial distribution of the local SNR . d) Relationship between the required sampling and the inverse local SNR . Each point represents one grid cell. e) Comparison between bootstrap estimates of the required sampling and the analytical prediction $N_{th} = \frac{r^2}{1-r^2} \frac{1}{SNR}$.

Do observations follow the same recoverability relationship?

A stringent independent test of the analytical scaling law is provided by observations, which represent a single realization of the climate system ($N=1$). In this limit, recoverability depends solely on local SNR, allowing the expected agreement with the forced response to be predicted without further calibration. Predicted recoverability differs substantially among precipitation, near-surface temperature and sea-level pressure (Fig. 3), reflecting their contrasting forced-to-internal variability ratios.

The theoretical predictions are confirmed across all three climate variables (Fig. 3). Observed agreement increases monotonically with the recoverability predicted for a single realization (Fig. 3c,f,i), with temperature exhibiting higher recoverability than precipitation and sea-level pressure, consistent with their contrasting SNR. Equivalently, observed agreement decreases as the sampling required to recover the forced response increases (Fig. 3b,e,h), demonstrating that regional model-observation agreement follows the same recoverability relationship derived from climate-model ensembles.

Previous studies have emphasized that observations cannot be expected to reproduce externally-forced regional climate responses exactly because they represent only one realization of the climate system, motivating the widespread use of large ensembles for model evaluation and uncertainty quantification^{11,17,31,32}. In seasonal and decadal prediction, SNR has likewise been recognized as a key determinant of predictive skill and model-observation agreement^{24–26,33}. Our results extend these ideas by showing that the expected level of regional model-observation agreement is itself quantitatively predictable from the local SNR through the same analytical scaling law that governs ensemble convergence.

The observational validation is robust across datasets and temporal scales. ERA5, CRU and GPCC all exhibit the same monotonic relationship between recoverability and observed agreement (Extended Data Fig. 10), indicating that the results are not sensitive to the observational product used to estimate the climate response. The same organization also persists after applying 10-year running means (Extended Data Fig. 11). Although temporal averaging suppresses a substantial fraction of interannual internal variability and thereby increases the local SNR, the analytical scaling law continues to describe the observed agreement. This behaviour is consistent with previous studies showing that averaging over decadal timescales enhances climate predictability while preserving an important role for internally-generated variability^{24–26,33}. Recoverability therefore appears to represent a general property of externally-forced regional climate responses that is robust across observational datasets, climate variables and temporal scales.

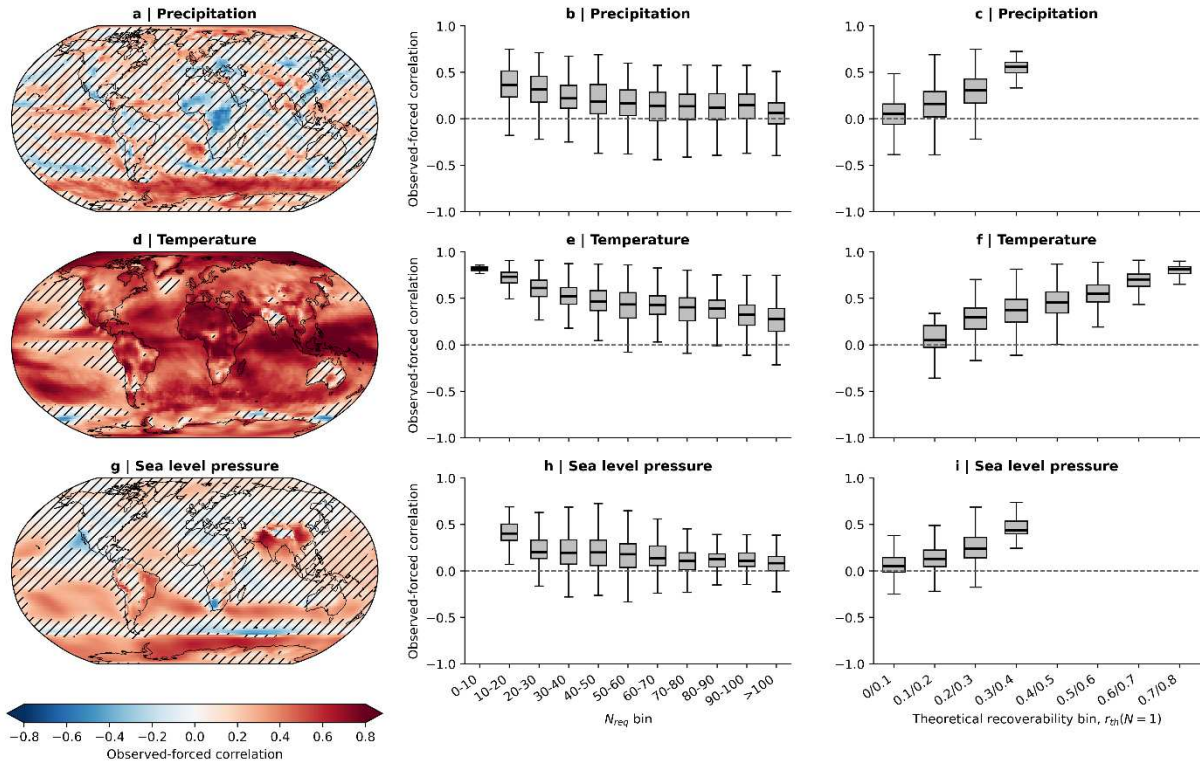


Figure 3 Regional model-observation agreement follows predicted climate recoverability. a, d, g, Temporal correlations between observed (ERA5) precipitation, near-surface temperature and mean sea-level pressure and independently estimated multi-model forced responses over 1940-2014. Hatching denotes correlations that are not statistically significant ($P > 0.05$). b, e, h, Distributions of observed-forced agreement as a function of the sampling required, N_{req} , to get the prescribed agreement $r=0.7$. Regions requiring greater sampling exhibit systematically weaker agreement between the observational realization and the forced response. c, f, i, Distribution of observed-forced correlations as a function of the theoretical recoverability $r = (1 + \frac{1}{N(SNR)})^{-1/2}$ expected from a single realization ($N=1$). Observed agreement increases monotonically with predicted recoverability across precipitation, near-surface temperature and mean sea-level pressure.

When does regional climate change become recoverable?

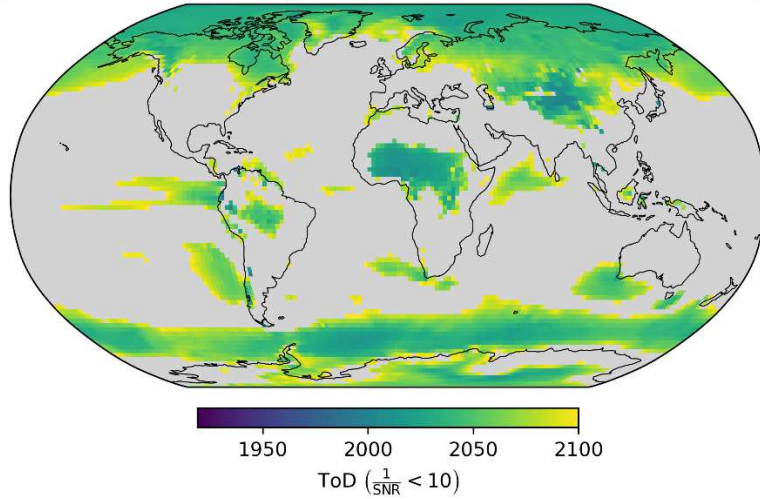
Climate emergence is commonly defined as the time at which an externally-forced response becomes distinguishable from internal variability, but its estimated timing depends strongly on the variable, region, reference period and signal-to-noise threshold considered^{6,20,22,23,34}. The recoverability framework proposed here provides a direct analytical interpretation of this dependence. For the single realization available from the observed climate system ($N=1$), the expected agreement with the forced response is $r_{(N=1)} = (1 + \frac{1}{SNR})^{-1/2}$, so that climate emergence corresponds to the temporal evolution of local SNR along the $N=1$ trajectory of the recoverability space (Fig. 2a). This formulation translates conventional emergence thresholds into explicit levels of expected agreement. For example, $1/SNR < 10$ corresponds to $r \approx 0.30$ for a single realization. At the same SNR, increasing the sampling to $N=10$ raises the expected agreement to approximately $r \approx 0.70$. A climate response that is only weakly recoverable from the observational record may therefore be recovered much more robustly from an ensemble, illustrating that emergence and ensemble convergence represent alternative trajectories through the same recoverability space.

We evaluate this interpretation using expanding historical and SSP2-4.5 simulations beginning in 1850. For each prescribed recoverability criterion (for example $1/\text{SNR} < 10$), we identify the first year at which the corresponding signal-to-noise threshold is satisfied (Fig. 4; Extended Data Figs. 12-13). Near-surface temperature becomes recoverable earliest and over the largest fraction of the globe. Large areas satisfy the $1/\text{SNR} < 10$ criterion during the twentieth century, and progressively more stringent recoverability levels expand across both land area and population through the twenty-first century (Extended Data Fig. 12). This early emergence is consistent with previous studies showing that anthropogenic warming rapidly exceeds internal temperature variability at regional scales^{21,23,34}. Precipitation emerges substantially later and much less uniformly (Fig. 4). The $1/\text{SNR} < 10$ criterion is reached first in regions where forced hydroclimatic responses are relatively strong, whereas extensive areas remain below this weak-recovery threshold by 2100. More stringent recoverability criteria are satisfied over only limited fractions of global land and population. This behaviour is consistent with earlier work showing that precipitation emergence is delayed by large internal variability and by the regional heterogeneity of forced hydrological and circulation responses^{10,20,35}.

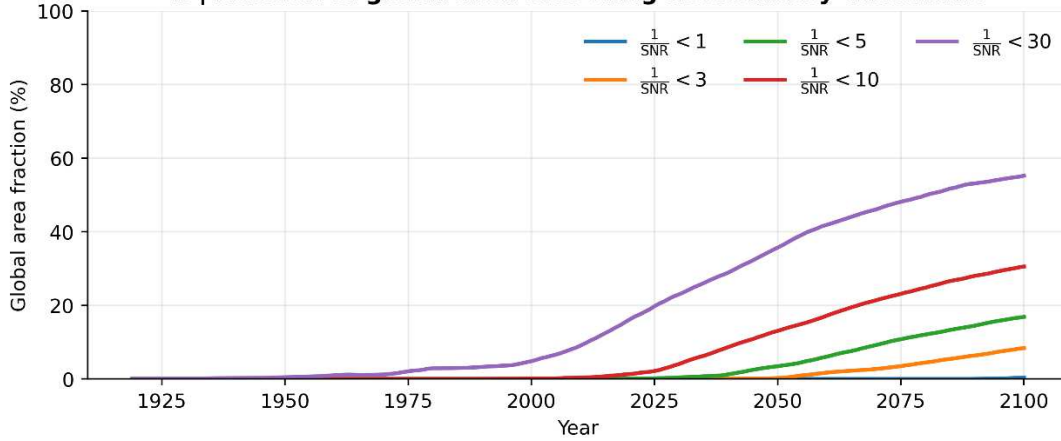
Sea-level pressure is even less recoverable over much of the globe (Extended Data Fig. 13). Large regions fail to satisfy even permissive criteria by 2100, and only a modest fraction of land area and population reaches the more stringent levels of agreement. This weak emergence is consistent with the dominant influence of atmospheric circulation variability on regional pressure responses and with the comparatively weak and spatially heterogeneous forced signal in circulation-related variables^{9,11}. The contrast among variables therefore reproduces the established time of emergence hierarchy: temperature emerges earliest, precipitation later, and circulation-related variables remain difficult to distinguish from internal variability. The recoverability framework adds a quantitative interpretation of this hierarchy by expressing every emergence criterion as a prescribed level of agreement. The large separation among the land-area and population curves in Fig. 4 and Extended Data Figs. 12-13 shows that regional climate emergence is not characterized by a unique date. More permissive criteria are satisfied decades before stringent ones, and the fraction of the population living in recoverable regions depends directly on the level of agreement required.

Previous studies have similarly shown that the global area and population exposed to climates outside historical variability increase progressively through time, with particularly early departure in low-latitude regions^{21,34}. Our results are broadly consistent with this pattern, but differ in interpretation: the curves represent not a binary transition into a new climate state, but a progression through increasingly stringent levels of recoverability. Regional climate emergence should therefore be viewed as a continuum whose timing depends jointly on the evolving SNR, the available sampling and the agreement required for interpretation.

a | Time of Detectability (ToD) of forced precipitation changes



b | Fraction of global area exceeding detectability thresholds



c | Fraction of global population living in detectable precipitation-change regions

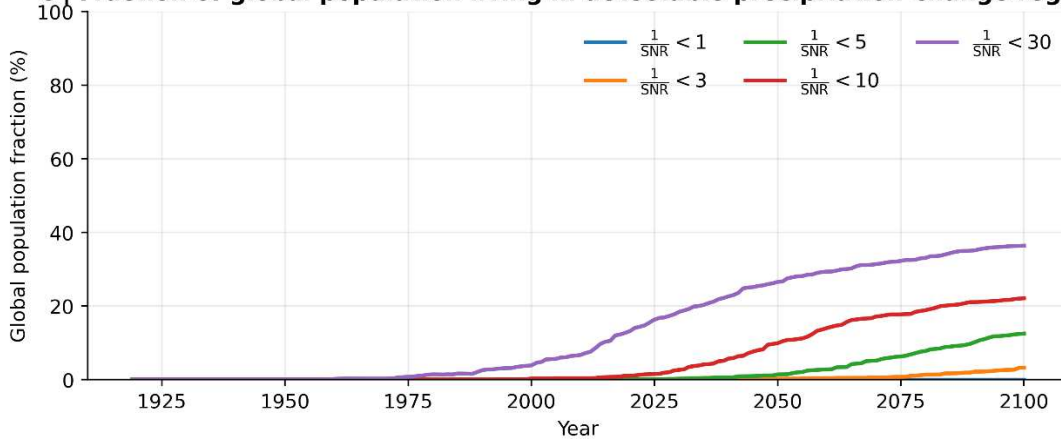


Figure 4 Regional climate emergence within the recoverability framework. a) First year at which forced precipitation responses satisfy the illustrative recoverability criterion $\frac{1}{SNR} < 10$, estimated from expanding historical and SSP2-4.5 simulations beginning in 1850. For a single realization, this threshold corresponds to an expected agreement of $r \approx 0.30$. Grey regions do not satisfy this criterion by 2100. b) Fraction of global land area satisfying progressively more stringent recoverability criteria. Thresholds of $\frac{1}{SNR} < 1, 3, 5, 10$ and 30 correspond approximately to $r \approx 0.7, 0.50, 0.4, 0.30$ and 0.18 , respectively, for $N=1$. c) Fraction of the global population living in regions satisfying the same criteria.

Discussion

Interpreting regional climate change remains challenging because externally forced responses are observed through internally generated variability^{3,6,18,29,31}. Here we show that problems spanning large-ensemble modelling, model evaluation and climate emergence can be understood within a common analytical framework linking local SNR, sampling and the agreement attainable with the underlying forced response. Rather than introducing another diagnostic metric, the framework predicts how much information about externally forced regional climate change can be recovered from finite sampling. The conceptual advance does not lie in recognizing that larger ensembles improve estimates of forced climate change or that SNR governs regional uncertainty; these principles are well established^{4,24–26}. Rather, we show that their interaction follows a predictable analytical scaling, placing ensemble convergence, model-observation agreement and climate emergence within the same quantitative framework.

An important implication concerns climate-model evaluation. Because observations represent a single realization ($N=1$), apparent discrepancies between observed and simulated regional responses should be interpreted in the context of finite sampling before being attributed to model deficiencies. This perspective complements efforts to account explicitly for internal variability in model evaluation^{17,36,37}, particularly for hydroclimatic variables where internal variability strongly masks the forced response^{10,12,38}. Regional differences in recoverability can also be related to their underlying forcing characteristics. Single-forcing experiments show that regions exhibiting a coherent response to a dominant external forcing generally require fewer realizations, whereas heterogeneous or competing forcing responses increase sampling requirements (Supplementary note and Figs. 1-2). This relationship provides a physical interpretation of regional SNR differences and their associated sampling requirements.

The same analytical relationship naturally extends to future climates. A key implication is that historical model-observation agreement should not provide equal confidence in future projections everywhere. Where the forced response is historically recoverable, observations provide an informative test of the simulated response underlying future projections; where recoverability is low, internal variability fundamentally limits such evaluation. The framework therefore provides a basis for distinguishing discrepancies that may indicate limitations in simulated forced responses from those expected from finite observational sampling. As local SNR evolves, it further predicts how these limits change through time (Extended Data Fig. 14), extending conventional time-of-emergence concepts^{21–23,34} from identifying when forced change becomes distinguishable to quantifying how robustly regional climate responses can be interpreted.

Several limitations nevertheless deserve consideration. Estimates of local SNR are affected by structural model uncertainty, biases in forced responses and internal variability, observational uncertainty, and limited multi-model independence^{8,39,40}. These factors primarily affect the inputs to the theory rather than the analytical relationship itself. More fundamentally, forced responses and internal variability need not be strictly independent because forcing can alter the amplitude and dynamics of internal variability⁴¹. Nevertheless, tests across large ensembles show that regional forced-state dependence of internal variance does not systematically increase departures from the analytical prediction (Extended Data Fig. 15), supporting the scaling as a robust first-order approximation for annual-mean responses.

Whether this robustness extends to climate extremes remains an open question. Forced changes in variance and distributional shape, together with circulation changes and land–atmosphere feedbacks, may complicate the separation of forced responses from internal variability^{42–44}. Extending recoverability to quantiles or extreme-value distribution parameters would test whether analogous scaling relationships govern the recoverability of climate extremes and other impact-relevant indicators.

More broadly, the framework provides a quantitative basis for interpreting regional climate information. By linking recoverability directly to local SNR and sampling, it identifies when historical observations meaningfully constrain simulated regional responses, where additional sampling can increase confidence, and where internal variability imposes an irreducible limit on evaluation. This distinction can help assess and communicate both the robustness and the limits of regional climate information for climate assessment, risk evaluation and adaptation planning.

Methods and data

Climate model simulations and observational datasets

We analysed historical, future and single-forcing climate simulations from the Coupled Model Intercomparison Project Phase 6 (CMIP6)⁴⁵. The historical ensemble comprises 782 realizations from 65 CMIP6 model configurations, including 100 members from the CESM2 Large Ensemble (CESM2-LE)^{46,47}, providing an extensive sampling of internally generated climate variability over 1940–2014. Detailed analyses of ensemble convergence focus on the ten Single Model Initial-condition Large Ensembles (SMILEs) containing more than 25 realizations, which allow robust estimation of externally forced responses independently within each model. Future analyses use 324 SSP2-4.5 realizations (2015–2100), while the physical drivers of recoverability are investigated using Detection and Attribution Model Intercomparison Project (DAMIP) single-forcing experiments, including greenhouse-gas (GHG), anthropogenic aerosol (AER) and natural (NAT) simulations. The complete list of climate models and ensemble members is provided in Extended Data Table 1.

Observed precipitation, near-surface air temperature and mean sea-level pressure were primarily obtained from the ERA5 reanalysis^{48,49}. To assess the robustness of precipitation results, analyses were repeated using the independent CRU TS⁵⁰ and GPCC⁵¹ observational datasets. Unless otherwise stated, all analyses use annual means over 1940–2014.

To ensure spatial consistency, all climate-model simulations and observational datasets were conservatively remapped onto a common $2^\circ \times 2^\circ$ latitude-longitude grid using the Climate Data Operators (CDO) remapcon interpolation scheme⁵². All subsequent analyses were performed on this common grid.

Estimation of externally-forced and internally-generated variability

The climate response at each grid point was decomposed into an externally forced component and an internally-generated component following the standard large-ensemble framework^{3,16}. For each model, the externally-forced response was estimated as the ensemble mean across all available realizations, while internally-generated variability was defined as the deviation of each individual realization from this ensemble mean, $X_i(t) = F(t) + I_i(t)$, where $X_i(t)$ is the

climate response of realization i , $F(t)$ is the externally forced response and $I_i(t)$ represents internally generated variability.

The local SNR was quantified as the ratio between the temporal variance of the externally-forced response and the temporal variance of internally-generated variability $SNR = \frac{\sigma_F^2}{\sigma_I^2}$, where σ_F^2 represents the temporal variance of the forced ensemble-mean response and σ_I^2 represents the mean temporal variance of internal variability across ensemble members. Regions with large SNR are characterized by externally-forced responses that dominate internal variability, whereas regions with low SNR remain strongly influenced by climate noise. This definition provides the statistical basis for all subsequent recoverability analyses.

Inter-model coherence analysis

To quantify the consistency of externally-forced climate responses across models, we adopted a leave-one-model-out framework. For each climate model, the externally-forced response was estimated from the ensemble mean of all available realizations and compared with an independent reference response constructed from the ensemble means of all remaining models. This procedure ensures complete statistical independence between the evaluated model and its reference.

Temporal Pearson correlations were computed independently at each grid point over the 1940-2014 period using annual means. Statistical significance was assessed using two-sided Pearson correlation tests ($P < 0.05$). For the probability density analyses (Fig. 1c and Extended Data Fig. 4), spatial distributions of grid-point correlations were estimated using kernel density estimation. Leave-one-model-out ensemble means were weighted according to the number of realizations contributed by each model to ensure that models with larger ensembles contributed proportionally to the multi-model reference.

Bootstrap estimation of recoverability

Regional recoverability was estimated independently at each grid point using non-parametric bootstrap resampling of the historical CMIP6 ensemble. For each ensemble size N , realizations were randomly sampled with replacement from the complete set of 782 historical simulations, and the corresponding ensemble mean was computed. Ensemble size was varied from 1 to 1000 members, and 1000 bootstrap samples were generated for each value of N . Because resampling is performed with replacement, N denotes bootstrap sample size rather than the number of distinct simulations. Values of N_{req} exceeding the 782 available historical realizations are therefore interpreted as lying beyond the empirically constrained sampling range rather than as precise ensemble-size estimates.

Each bootstrap ensemble mean was compared with a fixed reference forced response, defined as the ensemble mean of the complete historical ensemble, using the Pearson temporal correlation over 1940-2014. The resulting bootstrap distributions describe the level of agreement expected between a finite ensemble and the underlying externally-forced response as a function of available sampling.

Recoverability was quantified by the required ensemble size, N_{req} , defined as the smallest ensemble size for which the 5th percentile of the bootstrap correlation distribution exceeded a prescribed agreement threshold, $P_5[r(N)] \geq r_c$ where r_c denotes the target correlation. Unless

otherwise stated, $r_c = 0.7$, corresponding to robust recovery of the externally forced response. Sensitivity analyses using alternative correlation thresholds are presented in Extended Data Table 2 and show that the bootstrap estimates consistently follow the analytical recoverability relationship across a wide range of target agreements. Using the full historical ensemble as the reference forced response ensures that bootstrap resampling modifies only the amount of available sampling while leaving the target response unchanged. Recoverability therefore quantifies the information required to estimate a fixed externally forced climate signal rather than differences arising from changes in the reference response.

Theoretical relationship between signal-to-noise ratio and recoverability

Following the decomposition introduced above, the climate response of realization i at a given grid point is written as: $X_i(t) = F(t) + I_i(t)$, where $F(t)$ denotes the externally forced response and $I_i(t)$ the internally-generated component, assumed to have zero mean and to be approximately independent among realizations, $E[I_i(t)] \approx 0$.

The ensemble mean of a sufficiently large number of realizations therefore converges towards the externally-forced response, $\bar{X}(t) \approx F(t)$, which is treated throughout this study as the ensemble-mean response, which suppresses member-specific internal variability as ensemble size increases.

For an ensemble containing N independent realizations, $\bar{X}_N(t) = F(t) + \bar{I}_N(t)$, where $\bar{I}_N(t) = \frac{1}{N} \sum_{i=1}^N I_i(t)$

Assuming independence among realizations, ensemble averaging reduces the variance of internally generated variability according to $Var(\bar{I}_N) = \frac{\sigma_I^2}{N}$. The expected agreement between an N -member ensemble mean and the forced response can therefore be expressed as the Pearson correlation $r_N = Corr(\bar{X}_N, F)$. Assuming that the forced and internally-generated components are uncorrelated, $Cov(F, \bar{I}_N) \approx 0$, this relationship simplifies to $r_N = \frac{1}{\sqrt{1 + \frac{\sigma_I^2}{N\sigma_F^2}}}$ where: $\sigma_F^2 = Var(F)$

and $\sigma_I^2 = Var(I)$.

Defining the signal-to-noise ratio as $SNR = \frac{\sigma_F^2}{\sigma_I^2}$ yields the analytical recoverability relationship

$$r_N = \frac{1}{\sqrt{1 + \frac{1}{N(SNR)}}} \text{ which predicts the expected agreement attainable for a given level of sampling}$$

and local SNR.

Regional recoverability is quantified by the minimum sampling requirement, N_{opt} , defined as the smallest ensemble size for which the 5th percentile of the bootstrap correlation distribution exceeds a prescribed target agreement r_c . Approximating the bootstrap distribution by its expected value $r_N = r_c$ gives: $N_{th} = \frac{r_c^2}{1-r_c^2} \frac{1}{SNR}$, demonstrating that the sampling required to recover externally-forced regional climate responses is inversely proportional to the local SNR. This analytical relationship forms the theoretical basis of the recoverability framework and is evaluated against the bootstrap estimates throughout the study. The analytical expression predicts the ensemble size required for the expected recoverability to reach r_c , whereas our bootstrap definition of N_{req} is deliberately more conservative, requiring the 5th percentile of the

empirical recoverability distribution to exceed r_c . The two quantities are therefore not expected to coincide exactly.

Independence of forced and internal components

We decompose each realization as $X=F+I$, where F denotes the externally forced response and I internal variability. In general, $Var(X) = \sigma_F^2 + \sigma_I^2 + 2Cov(F, I)$. Our analytical recoverability relationship assumes approximate statistical independence between the forced and internally generated components, such that $Cov(F, I) \approx 0$, giving $Var(X) = \sigma_F^2 + \sigma_I^2$.

This approximation need not hold exactly if external forcing modifies the amplitude or structure of internal variability. We therefore evaluated its robustness using large initial-condition ensembles (Extended Data Fig. 15). Within each model, members were repeatedly divided into two independent subsets: one subset was used to estimate the forced response and the other to estimate internal variability and empirical recoverability, with the roles subsequently reversed. We quantified state dependence as $Corr[F, Var(I)]$ and tested whether its magnitude was associated with departures between empirical and theoretically predicted recoverability.

Because the forced response used for validation is itself estimated from a finite reference ensemble of N_{ref} members, its variance contains residual internal variability. Under the independence approximation, $Var(F_{ref}) = \sigma_F^2 + \frac{\sigma_I^2}{N_{ref}}$, such that we estimate $\sigma_F^2 = Var(F_{ref}) - \frac{\sigma_I^2}{N_{ref}}$

The recoverability expected against this finite-ensemble reference is therefore

$$r_{predicted} = \frac{\sigma_F^2}{\sqrt{(\sigma_F^2 + \sigma_I^2)(\sigma_F^2 + \sigma_I^2/N_{ref})}}$$

We compare this quantity with the independently cross-fitted empirical recoverability in Extended Data Fig. 15.

Observational evaluation of recoverability

Observed annual-mean precipitation, near-surface air temperature and mean sea-level pressure from ERA5 were compared with independently estimated multi-model forced responses using grid-point temporal Pearson correlations over 1940-2014. Precipitation analyses were repeated using the independent GPCC and CRU datasets to assess observational robustness. Statistical significance was evaluated using two-sided Pearson correlation tests ($P < 0.05$). Additional analyses were performed after applying 10-year running means to both observations and climate simulations.

Physical origins of regional climate recoverability

The physical drivers of recoverability were investigated using DAMIP single-forcing simulations. Ensemble-mean responses forced by greenhouse gases (GHG), anthropogenic aerosols (AER) and natural forcing (NAT) were compared with the corresponding historical all-forcing response using grid-point temporal Pearson correlations over 1940-2014. For each grid point, the dominant forcing was defined as the forcing experiment exhibiting the largest

statistically significant positive correlation with the historical response. Regions without significant positive correlations were left unclassified.

Recoverability through time

Time-dependent recoverability was evaluated using expanding historical and SSP2-4.5 simulations beginning in 1850. At each end year, the local SNR was recomputed using all years available from 1850 to that date. We use expanding windows beginning in 1850 because the objective is to quantify the progressive recoverability of the forced climate trajectory from the historical baseline, rather than recoverability of the contemporary climate state within a fixed moving window. The first year satisfying prescribed recoverability thresholds ($1/\text{SNR} < 1, 3, 5, 10$ and 30) was identified at every grid point. Global land-area fractions were computed as the percentage of land grid cells satisfying each recoverability criterion. Population exposure was then estimated by intersecting annual recoverability maps with annually varying ISIMIP historical and SSP2 population projections on the same $2^\circ \times 2^\circ$ grid.

Population exposure analysis

Population exposure was estimated using the ISIMIP bias-corrected historical and SSP2 annual population dataset⁵³, available through the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP; <https://www.isimip.org/gettingstarted/input-data-bias-adjustment/details/124/>). Annual population fields spanning 1850-2100 were conservatively remapped onto the common $2^\circ \times 2^\circ$ analysis grid using the same CDO remapcon interpolation applied to the climate data. For each year, the fraction of the global population living in regions satisfying a prescribed recoverability criterion was computed by summing the population within recoverable grid cells and dividing by the total global population for that year.

Statistical analyses

Unless otherwise stated, all correlations correspond to Pearson correlation coefficients computed from annual-mean time series. Statistical significance was assessed using two-sided tests ($P < 0.05$). Bootstrap confidence intervals correspond to the 5th and 95th percentiles of 1000 bootstrap realizations. All analyses were performed on the common $2^\circ \times 2^\circ$ grid using annual means unless explicitly stated otherwise.

Data availability

The climate model historical simulations, DAMIP and SSP2-4.5 projections are accessible through the Earth System Grid Federation (ESGF) archive as part of the Coupled Model Intercomparison Project Phase 6 (CMIP6) dataset at (<https://esgf-node.ipsl.upmc.fr/projects/cmip6-ipsl/>). ERA5 data is available from the ECMWF at <https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5>. CESM2 large ensemble data is available at <https://www.cesm.ucar.edu/community-projects/lens2/data-sets>. All data needed to evaluate and reproduce the results in the paper are publicly available.

Code availability

All analysis scripts used in this study will be freely available in GitHub.

Acknowledgements

This study received financial support from Région Nouvelle-Aquitaine project PSGAR GRIFON.

Contributions

R.A. conceived the study and developed the analytical framework. R.A., J.O. and H.D. designed the analyses. R.A. performed the analyses and produced the figures. R.A. wrote the initial manuscript. All authors discussed the results and reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Bindoff, N. L. *et al.* Detection and Attribution of Climate Change: from Global to Regional. in *Climate Change 2013: the Physical Science Basis* (2014).
2. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021). doi:10.1017/9781009157896.
3. Deser, C., Phillips, A., Bourdette, V. & Teng, H. Uncertainty in climate change projections: the role of internal variability. *Clim Dyn* **38**, 527–546 (2012).
4. Deser, C. “Certain Uncertainty: The Role of Internal Climate Variability in Projections of Regional Climate Change and Risk Management”. *Earth’s Future* **8**, e2020EF001854 (2020).
5. Lehner, F. *et al.* Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth System Dynamics* **11**, 491–508 (2020).
6. Time of emergence of climate signals - Hawkins - 2012 - Geophysical Research Letters - Wiley Online Library. <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2011gl050087>.
7. Simpson, I. R. *et al.* Confronting Earth System Model trends with observations. *Science Advances* **11**, eadt8035 (2025).
8. Knutti, R. & Sedláček, J. Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Clim Change* **3**, 369–373 (2013).
9. Shepherd, T. G. Atmospheric circulation as a source of uncertainty in climate change projections. *Nature Geosci* **7**, 703–708 (2014).

10. Douville, H., K. Raghavan, J. Renwick, R.P. Allan, P.A. Arias, M. Barlow, R. Cerezo-Mota, A. Cherchi, T.Y. Gan, J. Gergis, D. Jiang, A. Khan, W. Pokam Mba, D. Rosenfeld, J. Tierney, and O. Zolina, 2021: Water Cycle Changes. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1055–1210, doi: 10.1017/9781009157896.010. in.
11. Shaw, T. A. *et al.* Regional climate change: consensus, discrepancies, and ways forward. *Front. Clim.* **6**, (2024).
12. Gu, L. *et al.* Uncertain dynamic response of mid-latitude winter precipitation. *Nature* **653**, 110–116 (2026).
13. Kay, J. E. *et al.* The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. *Bulletin of the American Meteorological Society* **96**, 1333–1349 (2015).
14. Milinski, S., Maher, N. & Olonscheck, D. How large does a large ensemble need to be? *Earth System Dynamics* **11**, 885–901 (2020).
15. Mankin, J. S., Lehner, F., Coats, S. & McKinnon, K. A. The Value of Initial Condition Large Ensembles to Robust Adaptation Decision-Making. *Earth's Future* **8**, e2012EF001610 (2020).
16. Maher, N. *et al.* The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability. *Journal of Advances in Modeling Earth Systems* **11**, 2050–2069 (2019).
17. Suarez-Gutierrez, L., Milinski, S. & Maher, N. Exploiting large ensembles for a better yet simpler climate model evaluation. *Clim Dyn* **57**, 2557–2580 (2021).
18. Tett, S. F. B., Stott, P. A., Allen, M. R., Ingram, W. J. & Mitchell, J. F. B. Causes of twentieth-century temperature change near the Earth's surface. *Nature* **399**, 569–572 (1999).

- 555 19. Gillett, N. P., Zwiers, F. W., Weaver, A. J. & Stott, P. A. Detection of human influence on sea-
556 level pressure. *Nature* **422**, 292–294 (2003).
- 557 20. Giorgi, F. & Bi, X. Time of emergence (TOE) of GHG-forced precipitation change hot-spots.
558 *Geophysical Research Letters* **36**, (2009).
- 559 21. Hawkins, E. *et al.* Observed Emergence of the Climate Change Signal: From the Familiar to the
560 Unknown. *Geophysical Research Letters* **47**, e2019GL086259 (2020).
- 561 22. Douglas, H. C. *et al.* Changes to population-based emergence of climate change from CMIP5 to
562 CMIP6. *Environ. Res. Lett.* **18**, 014013 (2022).
- 563 23. King, A. D., Douglas, H., Harrington, L. J., Hawkins, E. & Borowiak, A. R. Climate change
564 emergence over people’s lifetimes. *Environ. Res.: Climate* **2**, 041002 (2023).
- 565 24. Smith, D. M. *et al.* Improved Surface Temperature Prediction for the Coming Decade from a
566 Global Climate Model. *Science* **317**, 796–799 (2007).
- 567 25. Scaife, A. A. & Smith, D. A signal-to-noise paradox in climate science. *npj Clim Atmos Sci* **1**, 1–8
568 (2018).
- 569 26. Eade, R. *et al.* Do seasonal-to-decadal climate predictions underestimate the predictability of
570 the real world? *Geophysical Research Letters* **41**, 5620–5628 (2014).
- 571 27. Boer, G. J. *et al.* The Decadal Climate Prediction Project (DCPP) contribution to CMIP6.
572 *Geoscientific Model Development* **9**, 3751–3777 (2016).
- 573 28. Weisheimer, A. *et al.* The Signal-to-Noise Paradox in Climate Forecasts: Revisiting Our
574 Understanding and Identifying Future Priorities. *Bulletin of the American Meteorological Society*
575 **105**, E651–E659 (2024).
- 576 29. Maher, N., Milinski, S. & Ludwig, R. Large ensemble climate model simulations: introduction,
577 overview, and future prospects for utilising multiple types of large ensemble. *Earth System*
578 *Dynamics* **12**, 401–418 (2021).
- 579 30. Hawkins, E. & Sutton, R. The Potential to Narrow Uncertainty in Regional Climate Predictions.
580 *Bulletin of the American Meteorological Society* **90**, 1095–1108 (2009).

- 581 31. Eyring, V. *et al.* Taking climate model evaluation to the next level. *Nature Clim Change* **9**, 102–
582 110 (2019).
- 583 32. Jain, S. *et al.* Importance of internal variability for climate model assessment. *npj Clim Atmos Sci*
584 **6**, 68 (2023).
- 585 33. Smith, D. M. *et al.* North Atlantic climate far more predictable than models imply. *Nature* **583**,
586 796–800 (2020).
- 587 34. Mora, C. *et al.* The projected timing of climate departure from recent variability. *Nature* **502**,
588 183–187 (2013).
- 589 35. Chadwick, R., Good, P., Martin, G. & Rowell, D. P. Large rainfall changes consistently projected
590 over substantial areas of tropical land. *Nature Clim Change* **6**, 177–181 (2016).
- 591 36. Suarez-Gutierrez, L., Li, C., Thorne, P. W. & Marotzke, J. Internal variability in simulated and
592 observed tropical tropospheric temperature trends. *Geophys Res Lett* **44**, (2017).
- 593 37. Suarez-Gutierrez, L., Maher, N. & Milinski, S. Evaluating the internal variability and forced
594 response in large ensembles. *CLIVAR Var* **18**, (2020).
- 595 38. Knutson, T. R. & Zeng, F. Model assessment of observed precipitation trends over land regions:
596 detectable human influences and Possible low bias in model trends. *J. Clim.* **31**, (2018).
- 597 39. Sanderson, B. M., Knutti, R. & Caldwell, P. Addressing interdependency in a multimodel
598 ensemble by interpolation of model properties. *J. Clim.* **28**, 5150–70 (2015).
- 599 40. Brunner, L. *et al.* Reduced global warming from CMIP6 projections when weighting models by
600 performance and independence. *Earth System Dynamics* **11**, 995–1012 (2020).
- 601 41. Lehner, F. & Deser, C. Origin, importance, and predictive limits of internal climate variability.
602 *Environ. Res.: Climate* **2**, 023001 (2023).
- 603 42. Seneviratne, S. I. *et al.* in *Climate Change 2021: The Physical Science Basis* (eds Masson-
604 Delmotte, V. *et al.*) Ch. 11 (Cambridge Univ. Press, 2021).
- 605 43. Vautard, R. *et al.* Heat extremes in Western Europe increasing faster than simulated due to
606 atmospheric circulation trends. *Nat Commun* **14**, 6803 (2023).

44. Rahmati, M. *et al.* Soil Moisture Memory: State-Of-The-Art and the Way Forward. *Reviews of Geophysics* **62**, e2023RG000828 (2024).
45. Eyring, V. *et al.* Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
46. Danabasoglu, G. *et al.* The Community Earth System Model Version 2 (CESM2). *Journal of Advances in Modeling Earth Systems* **12**, e2019MS001916 (2020).
47. Rodgers, K. B. *et al.* Ubiquity of human-induced changes in climate variability. *Earth System Dynamics* **12**, 1393–1411 (2021).
48. Bell, B. *et al.* The ERA5 global reanalysis: Preliminary extension to 1950. *Quarterly Journal of the Royal Meteorological Society* **147**, 4186–4227 (2021).
49. Hersbach, H. *et al.* The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146**, 1999–2049 (2020).
50. Harris, I., Osborn, T. J., Jones, P. & Lister, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci Data* **7**, 109 (2020).
51. Schneider, U. *et al.* GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theor Appl Climatol* **115**, 15–40 (2014).
52. Schulzweida, U. CDO User Guide. <https://doi.org/10.5281/zenodo.10020800> (2023)
doi:10.5281/zenodo.10020800.
53. Frieler, K. *et al.* Assessing the impacts of 1.5 °C global warming – simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b). *Geoscientific Model Development* **10**, 4321–4345 (2017).

Extended Data

Table 1 Climate models and ensemble sizes included in this study. Shown are the CMIP6 and CESM2 Large Ensemble (CESM2-LE) models contributing to the historical, SSP2-4.5, and DAMIP experiments. Models are listed alphabetically, and ensemble sizes indicate the number of realizations available for analysis. The ten largest ensembles are shown in bold.

Model Name	Ensemble size				
	historical	SSP2-45	GHG	AER	NAT
ACCESS-CM2	3	3			
ACCESS-ESM1-5	40	40	3	3	3
AWI-CM-1-1-MR	5	1			
AWI-ESM-1-1-LR	1				
BCC-CSM2-MR	3	1	3	3	3
BCC-ESM1	3				
CAMS-CSM1-0	3	2			
CAS-ESM2-0	4	2			
CESM2	10	3	3	2	3
CESM2-FV2	3				
CESM2-LE	100	3	15	20	
CESM2-WACCM	3				
CESM2-WACCM-FV2	3				
CIESM	3	1			
CMCC-CM2-HR4	1				
CMCC-CM2-SR5	2	1			
CMCC-ESM2	1	1			
CNRM-CM6-1	30	10			
CNRM-CM6-1-HR	1	1			
CNRM-ESM2-1	12	10			
CanESM5	65	50	50	30	50
CanESM5-CanOE	3	3			
E3SM-1-0	23		5	5	5
E3SM-1-1	1	1			
E3SM-1-1-ECA	1				
E3SM-2-0	21				
E3SM-2-0-NARRM	5				
E3SM-2-1	5				
EC-Earth3	14	14			
EC-Earth3-AerChem	2				
EC-Earth3-CC	1	1			
EC-Earth3-Veg	7	5			
EC-Earth3-Veg-LR	3	3			
FGOALS-f3-L	3	1			
FGOALS-g3	6	4	2	3	2

FIO-ESM-2-0	3	3			
GFDL-CM4	1	1			
GFDL-ESM4	3	3	1	1	3
GISS-E2-1-G	65	19		16	
GISS-E2-1-G-CC	1				
GISS-E2-1-H	25	5			
GISS-E2-2-G	11	5			
GISS-E2-2-H	5				
GISS-E3-G	2				
HadGEM3-GC31-LL	5	4			
HadGEM3-GC31-MM	4				
IITM-ESM	1	1			
INM-CM4-8	1	1			
INM-CM5-0	10	1			
IPSL-CM5A2-INCA	1				
IPSL-CM6A-LR	33	11			
IPSL-CM6A-LR-INCA	1				
KACE-1-0-G	3	3			
MIROC-ES2H	1				
MIROC-ES2L	29	20			
MIROC6	50	50			
MPI-ESM-1-2-HAM	3				
MPI-ESM1-2-HR	10				
MPI-ESM1-2-LR	50	10			
MRI-ESM2-0	12	5	5	5	5
NESM3	5	2			
NorCPM1	30				
NorESM2-LM	3	3	3	3	3
NorESM2-MM	1	2			
SAM0-UNICON	1				
TaiESM1	2	1			
UKESM1-0-LL	19	14			
Total	782	324	93	91	80

Table 2 Comparison of theoretical and empirical (bootstrap) scaling coefficients. Shown are the theoretical coefficient ($r_c^2/(1-r_c^2)$) and the slope estimated from the linear regression between the inverse signal-to-noise ratio ($1/\text{SNR}$) and the required ensemble size (N_{req}). The agreement between the two coefficients provides an assessment of the theoretical prediction.

Correlation threshold r_c	Theoretical coefficient $r_c^2/(1-r_c^2)$	Estimated slope
0.7	0.96	1.18
0.75	1.29	1.52
0.8	1.78	2.04

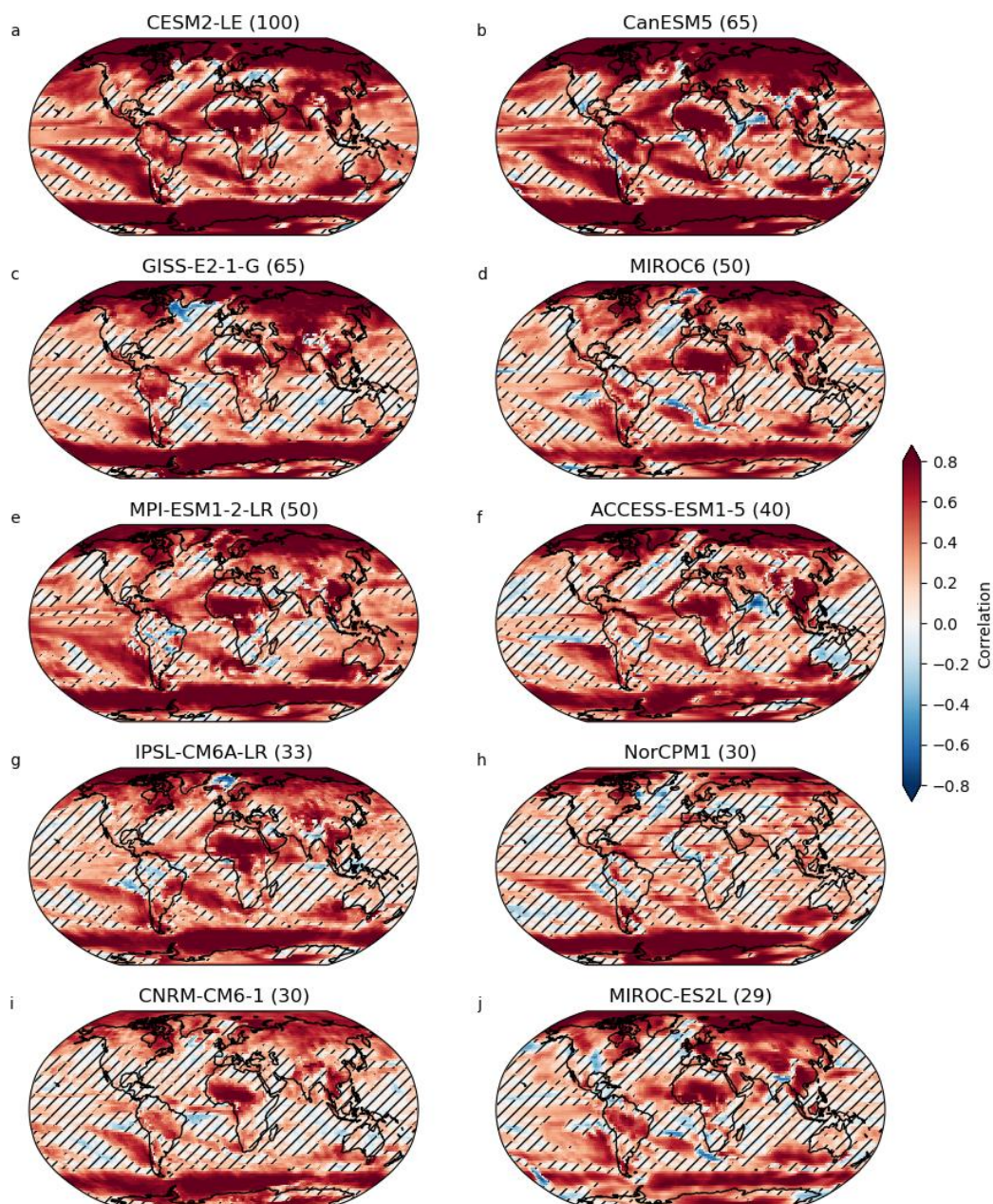


Figure 1 Inter-model coherence of forced precipitation responses. Spatial correlation between the ensemble-mean precipitation response of each model and that of all remaining models (leave-one-model-out framework) over 1940-2014. Correlations are computed at each grid point using annual means. Hatched regions indicate correlations not statistically significant at the 5% level. High correlations indicate robust externally forced responses across models.

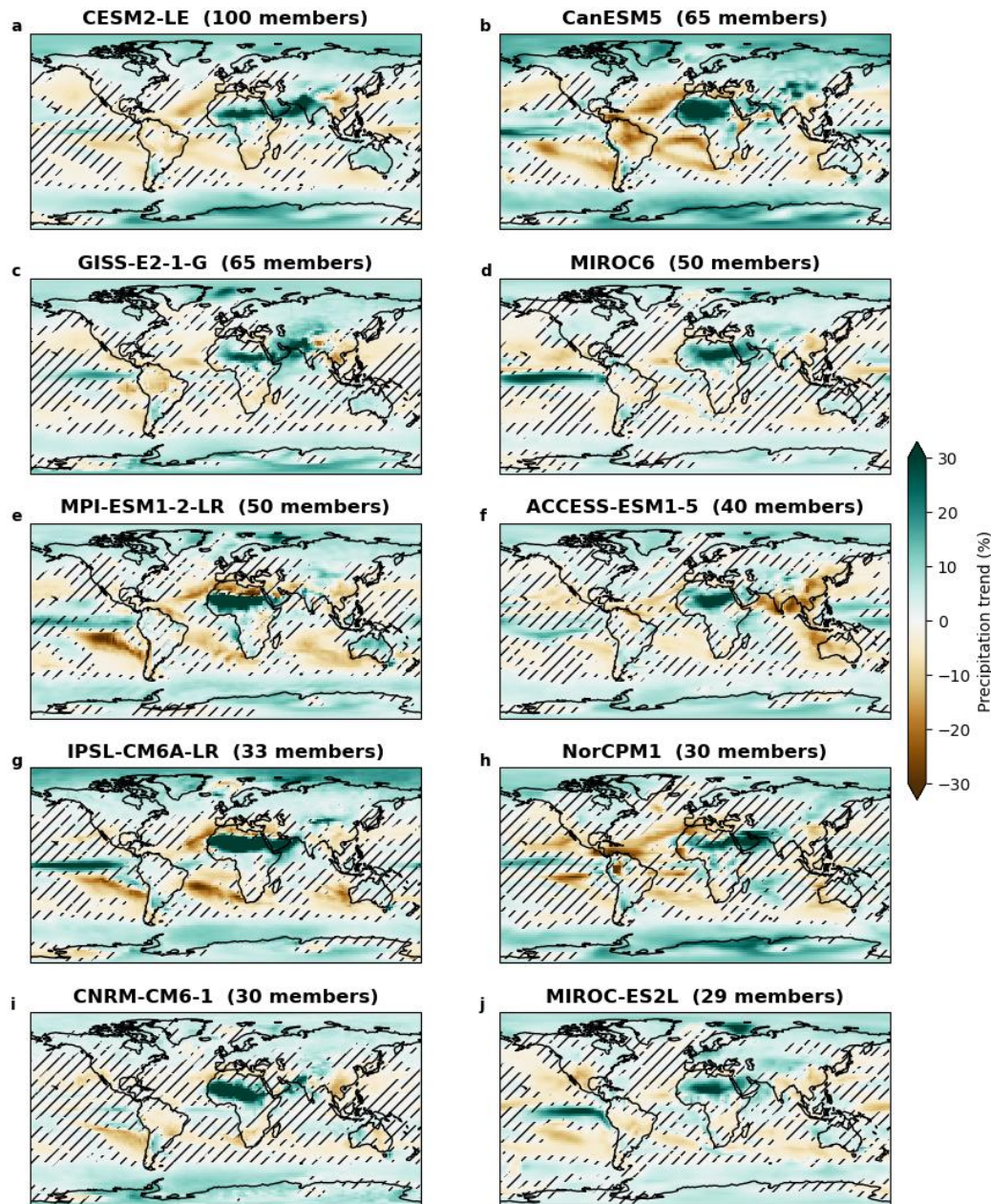


Figure 2 Ensemble-mean annual precipitation trends (in %) across the ten largest climate model ensembles. Each panel shows the linear trend in ensemble-mean annual precipitation over the period 1940-2014 for one of the ten largest available climate model ensembles included in this study. Trends are computed at each grid point using annual mean precipitation and are estimated from the ensemble mean to emphasize the externally forced response while reducing the influence of internal variability. Hatched regions indicate trends that are not statistically significant at the 5% level.

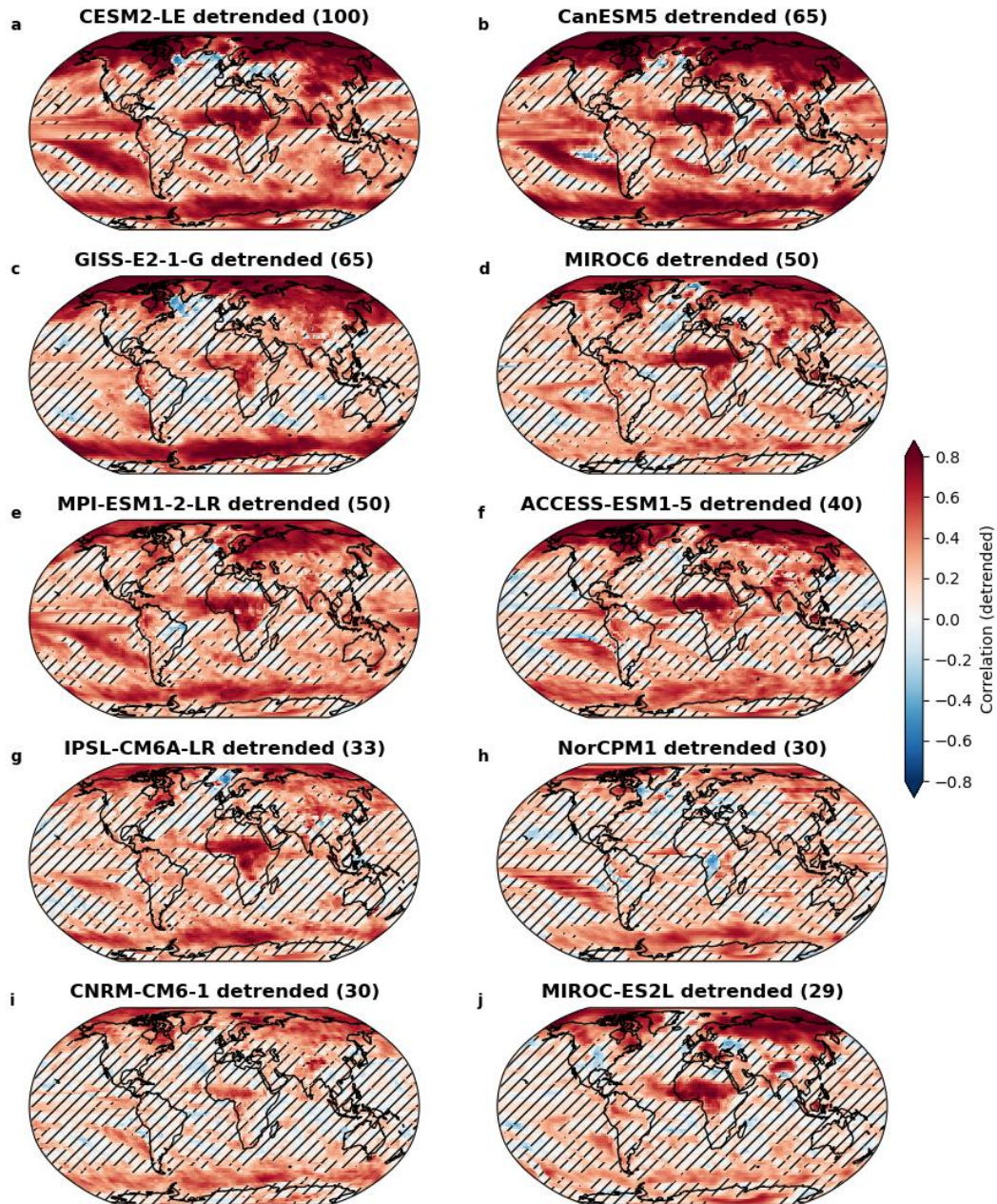


Figure 3 Inter-model coherence of forced detrended precipitation responses. Spatial correlation between the detrended ensemble-mean precipitation response of each model and that of all remaining models (leave-one-model-out framework) over 1940-2014. Correlations are computed at each grid point using annual means. Hatched regions indicate correlations not statistically significant at the 5% level. High correlations indicate robust externally forced responses across models.

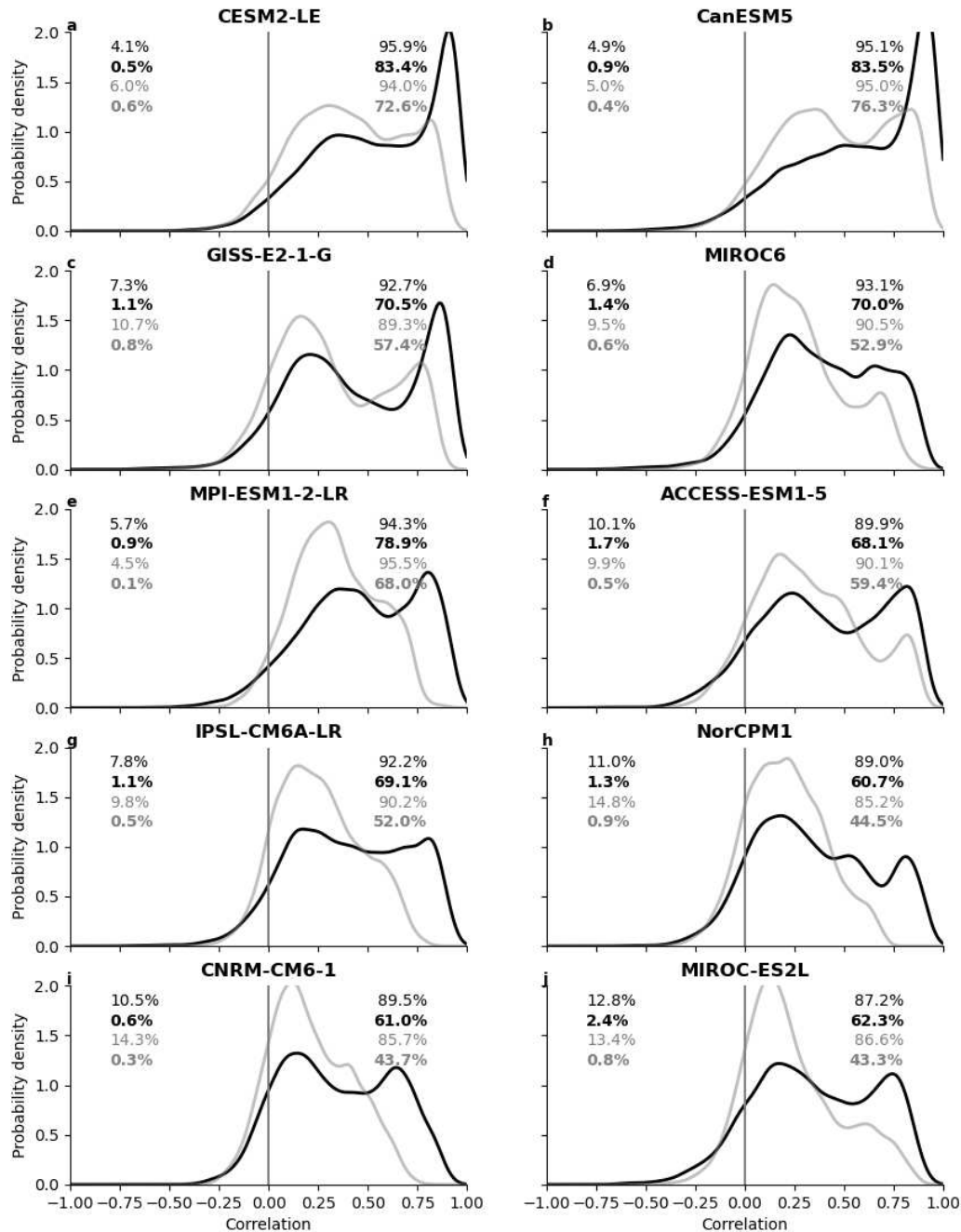


Figure 4 Probability density functions of spatial correlations estimated using kernel density estimation. Each curve represents the distribution of grid-point correlations for a given model relative to the ensemble mean of the remaining models. Percentages indicate the fraction of grid points with positive (right) and negative (left) correlations. Bold values denote the fraction of correlations that are statistically significant at the 5% level. The x-axis shows the Pearson correlation coefficient, and the y-axis represents probability density. Grey curves are same as black but for linearly detrended precipitation.

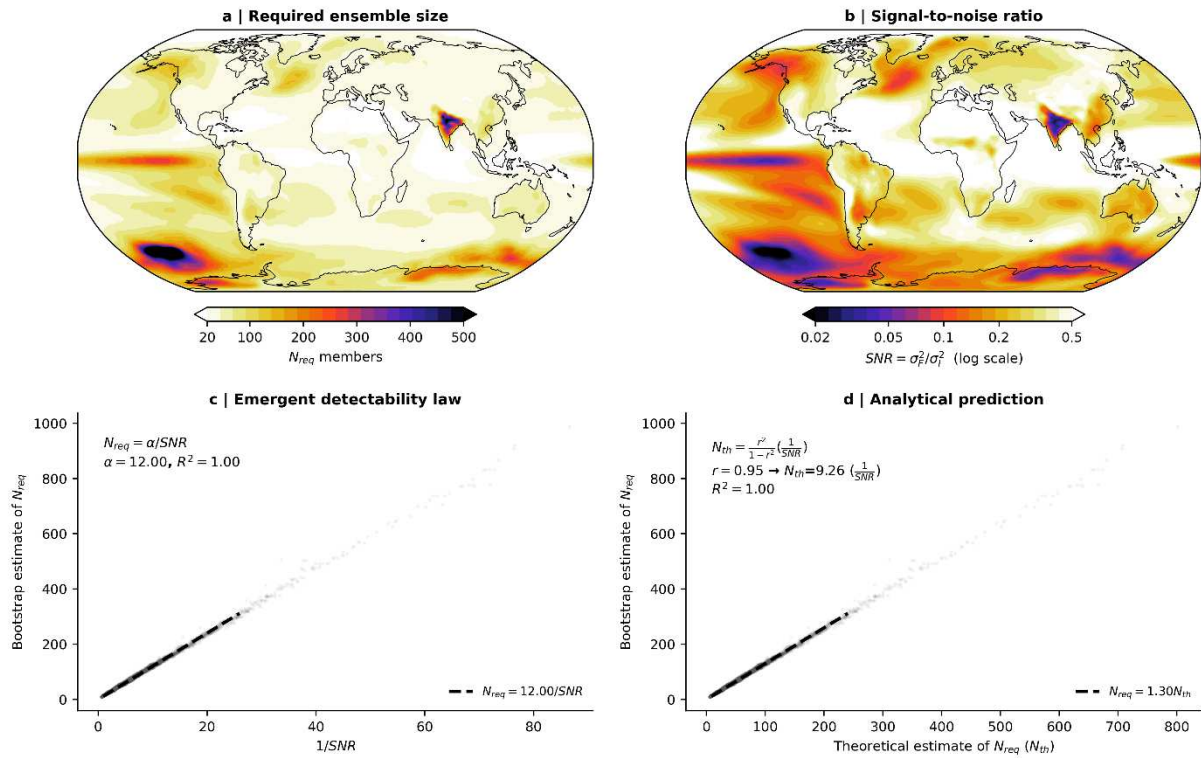


Figure 5 The signal-to-noise relationship governing regional climate recoverability is robust across climate variables. a) Spatial distribution of the required ensemble size (N_{req}) required for robust recovery of externally forced surface air temperature change over 1940-2014. b) Signal-to-noise ratio (SNR) computed from externally forced and internally generated air temperature variability. Low-SNR regions closely coincide with large ensemble-size requirements. c) Relationship between N_{req} and the inverse SNR. The dashed line shows the best-fit emergent scaling law $N_{req} = \alpha/SNR$. d) Comparison between bootstrap-derived ensemble-size estimates (N_{req}) and the analytical prediction (N_{th}). The dashed line indicates the best-fit proportionality relationship.

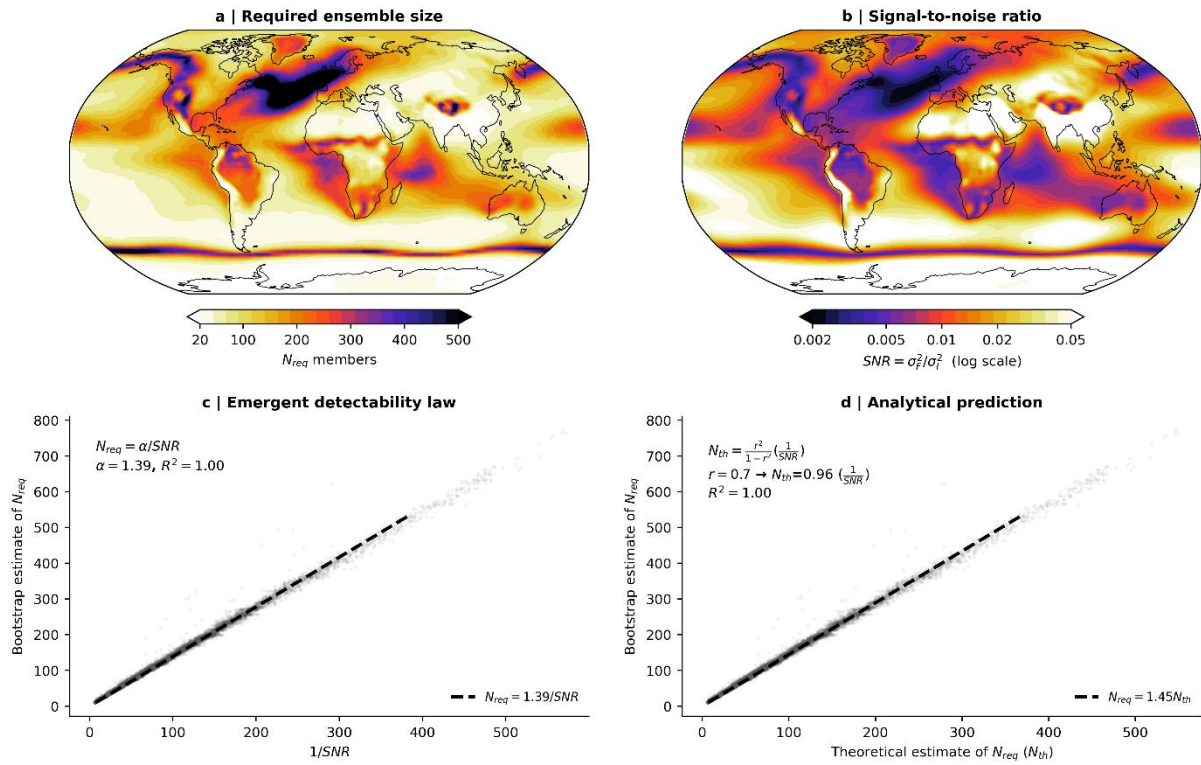


Figure 6 The signal-to-noise relationship governing regional climate recoverability is robust across climate variables. a) Spatial distribution of the required ensemble size (N_{req}) required for robust recovery of externally forced sea-level pressure change over 1940-2014. b) Signal-to-noise ratio (SNR) computed from externally forced and internally generated sea-level pressure variability. Low-SNR regions closely coincide with large ensemble-size requirements. c) Relationship between N_{req} and the inverse SNR. The dashed line shows the best-fit emergent scaling law $N_{req} = \alpha / \text{SNR}$. d) Comparison between bootstrap-derived ensemble-size estimates (N_{req}) and the analytical prediction (N_{th}). The dashed line indicates the best-fit proportionality relationship.

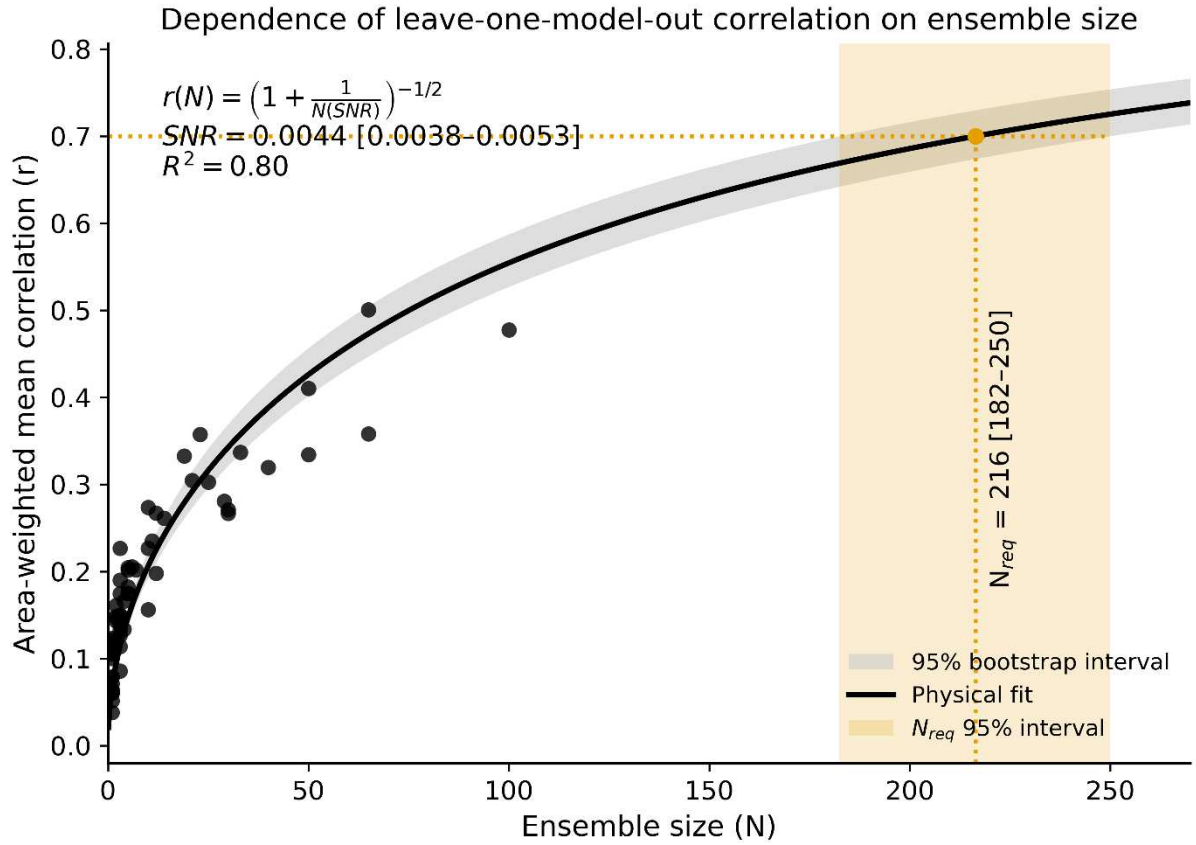


Figure 7 Inter-model agreement follows the theoretical signal-to-noise scaling. Area-weighted mean spatial correlation between the ensemble-mean precipitation response of individual CMIP6 models and an independent leave-one-model-out multi-model reference as a function of ensemble size (N). For each model, the forced response is estimated from the ensemble mean of that model and compared with a reference response constructed from all remaining models, ensuring statistical independence between the evaluated model and the reference field. Black circles show individual models, while the solid black curve indicates the best-fit signal-to-noise relationship, $r_N = \frac{1}{\sqrt{1 + \frac{1}{N(SNR)}}}$. Grey shading denotes the 95% bootstrap confidence interval obtained by resampling models. The horizontal dotted line marks the detectability threshold ($r=0.7$), and the vertical dotted line indicates the corresponding required ensemble size (N_{req}). Orange shading shows the bootstrap-derived 95% confidence interval of N_{req} .

The close agreement between inter-model correlations and the theoretical signal-to-noise prediction demonstrates that convergence toward a common forced hydroclimate response is strongly constrained by ensemble size and SNR. The larger ensemble-size requirement obtained from this leave-one-model-out framework relative to Fig. 1d likely reflects methodological differences between the two approaches. Whereas Fig. 1d estimates detectability locally from grid-point signal-to-noise characteristics before spatial aggregation, the leave-one-model-out analysis evaluates convergence using a single area-weighted measure of global inter-model agreement. The fitted parameter therefore represents an effective large-scale SNR rather than the spatially varying detectability diagnosed at individual grid points. Despite these differences, both approaches recover the same theoretical signal-to-noise dependence of ensemble agreement, supporting the robustness of the detectability framework.

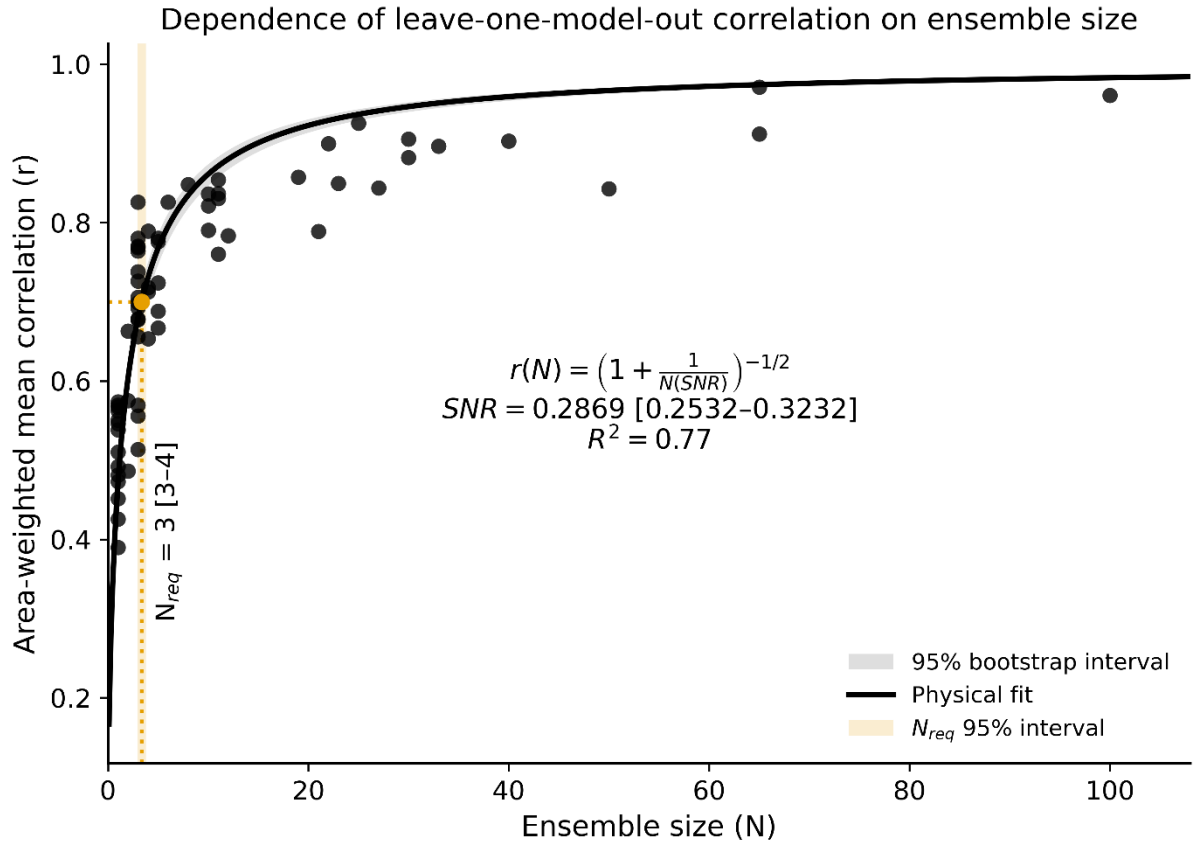


Figure 8 Inter-model agreement follows the theoretical signal-to-noise scaling. Area-weighted mean spatial correlation between the ensemble-mean surface air temperature response of individual CMIP6 models and an independent leave-one-model-out multi-model reference as a function of ensemble size (N). For each model, the forced response is estimated from the ensemble mean of that model and compared with a reference response constructed from all remaining models, ensuring statistical independence between the evaluated model and the reference field. Black circles show individual models, while the solid black curve indicates the best-fit signal-to-noise relationship, $r_N = \frac{1}{\sqrt{1 + \frac{1}{N(SNR)}}}$.

Grey shading denotes the 95% bootstrap confidence interval obtained by resampling models. The horizontal dotted line marks the detectability threshold ($r=0.7$), and the vertical dotted line indicates the corresponding required ensemble size (N_{req}). Orange shading shows the bootstrap-derived 95% confidence interval of N_{req} .

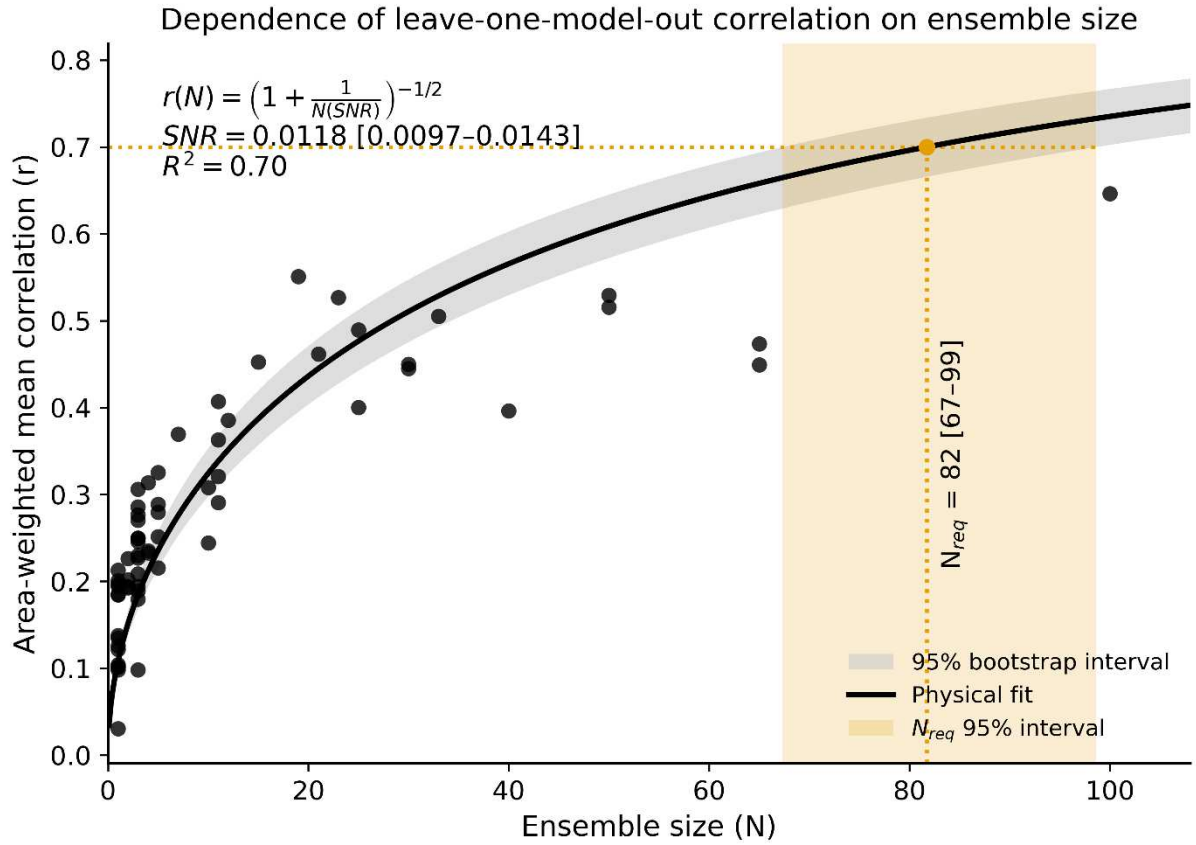


Figure 9 Inter-model agreement follows the theoretical signal-to-noise scaling. Area-weighted mean spatial correlation between the ensemble-mean sea-level pressure response of individual CMIP6 models and an independent leave-one-model-out multi-model reference as a function of ensemble size (N). For each model, the forced response is estimated from the ensemble mean of that model and compared with a reference response constructed from all remaining models, ensuring statistical independence between the evaluated model and the reference field. Black circles show individual models, while the solid black curve indicates the best-fit signal-to-noise relationship, $r_N = \frac{1}{\sqrt{1 + \frac{1}{N(SNR)}}}$. Grey shading denotes the 95% bootstrap confidence interval obtained by resampling models. The horizontal dotted line marks the detectability threshold ($r=0.7$), and the vertical dotted line indicates the corresponding required ensemble size (N_{req}). Orange shading shows the bootstrap-derived 95% confidence interval of N_{req} .

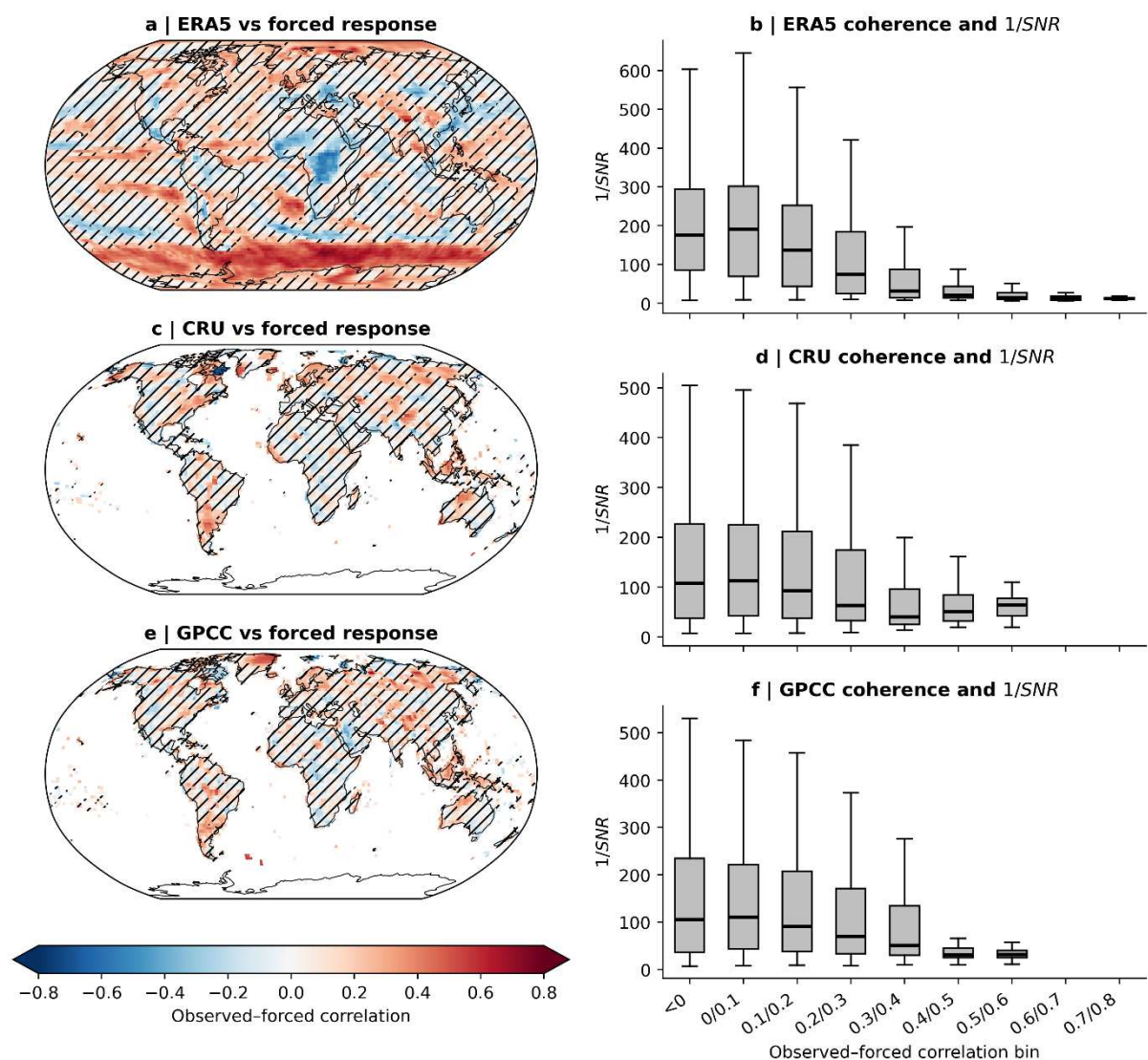


Figure 10 Independent observational datasets support the regional climate recoverability framework. a,c,e, Temporal correlations between observed precipitation products (ERA5, CRU and GPCC) and the full multi-model forced response. Hatching denotes non-significant correlations ($P > 0.05$). b,d,f, Distribution of the inverse local signal-to-noise ratio ($1/\text{SNR}$) across bins of observed–forced correlation. Regions exhibiting stronger observed coherence systematically exhibit lower inverse SNR, and therefore lower theoretical sampling requirements for recovering the forced response.

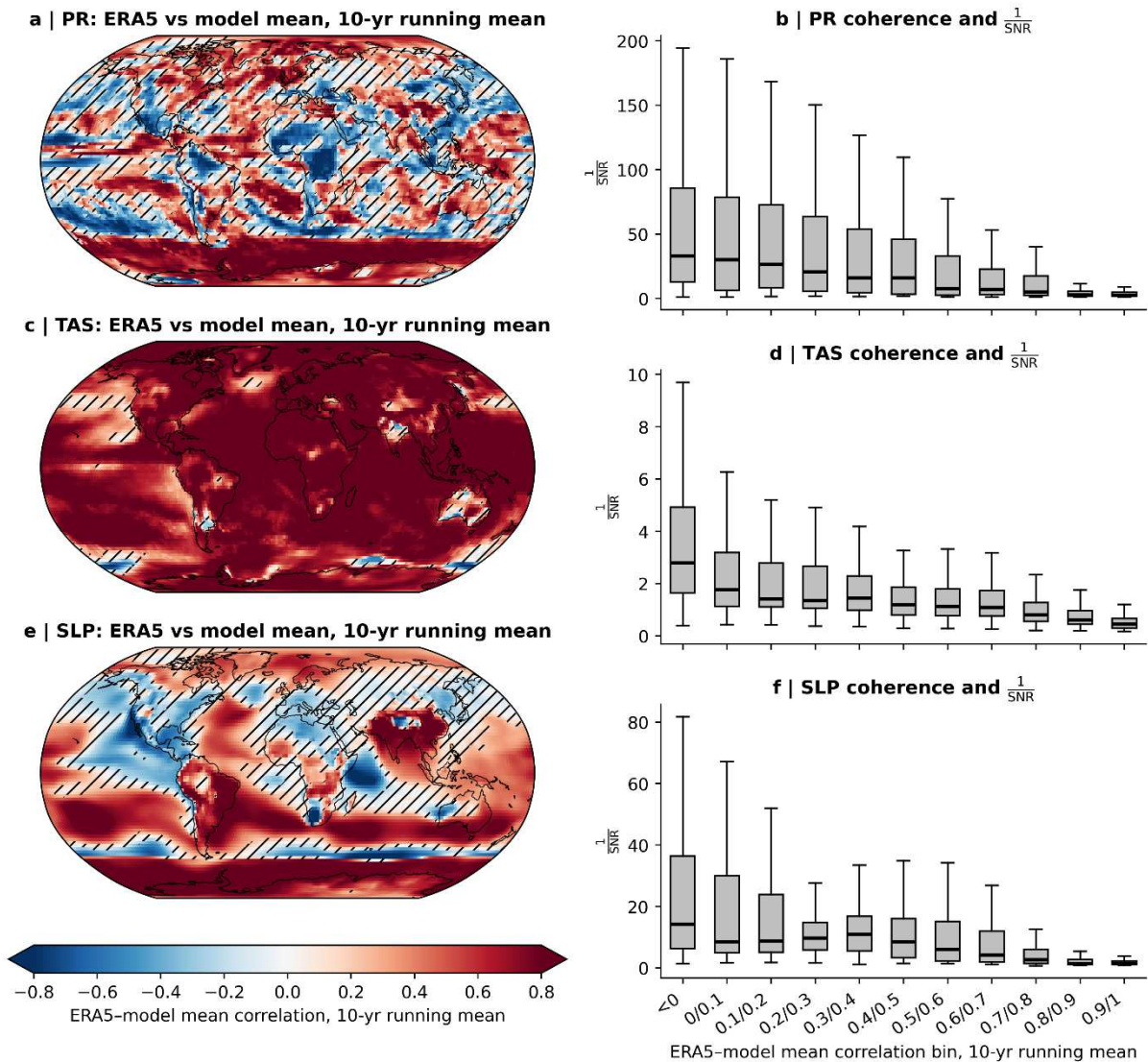


Figure 11 Regional climate recoverability is preserved across annual and decadal timescales. (a,c,e) Spatial correlation between ERA5 and the CMIP6 multi-model forced response for precipitation (PR), near-surface air temperature (TAS), and sea-level pressure (SLP), respectively, after applying a 10-year running mean to both observations and model simulations over 1940–2014. Hatched regions denote correlations that are not statistically significant ($P > 0.05$). (b,d,f) Distribution of the inverse local signal-to-noise ratio ($1/\text{SNR}$) for grid points grouped according to the corresponding ERA5-model mean correlation shown in panels (a,c,e). Correlation bins have a width of 0.1.

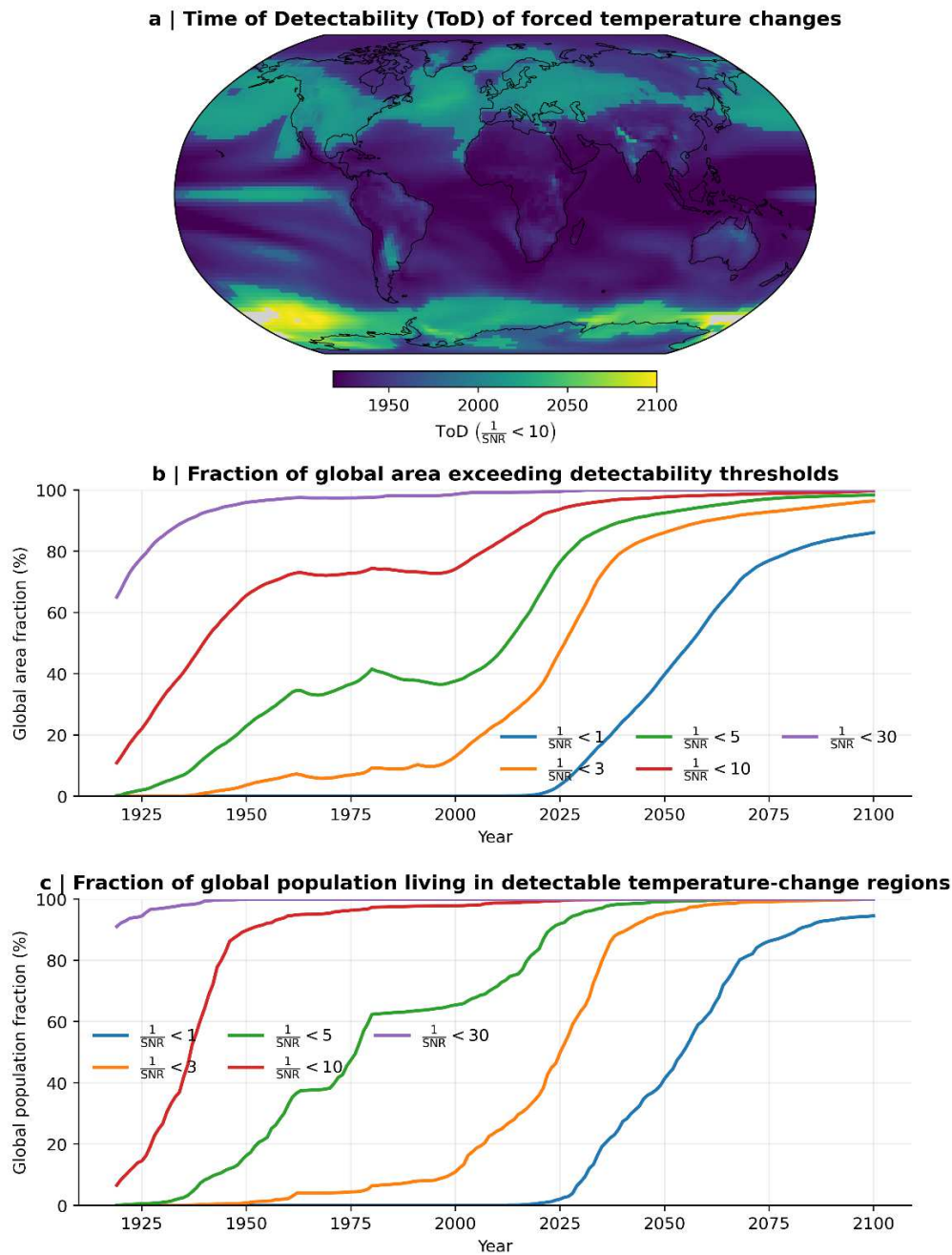


Figure 12 Regional climate emergence within the recoverability framework. a) First year at which forced temperature responses satisfy the illustrative recoverability criterion $1/\text{SNR} < 10$, estimated from expanding historical and SSP2-4.5 simulations beginning in 1850. For a single realization, this threshold corresponds to an expected agreement of $r \approx 0.30$. Grey regions do not satisfy this criterion by 2100. b) Fraction of global land area satisfying progressively more stringent recoverability criteria. Thresholds of $1/\text{SNR} < 1, 3, 5, 10$ and 30 correspond approximately to $r \approx 0.7, 0.50, 0.4, 0.30$ and 0.18 , respectively, for $N=1$. c) Fraction of the global population living in regions satisfying the same criteria.

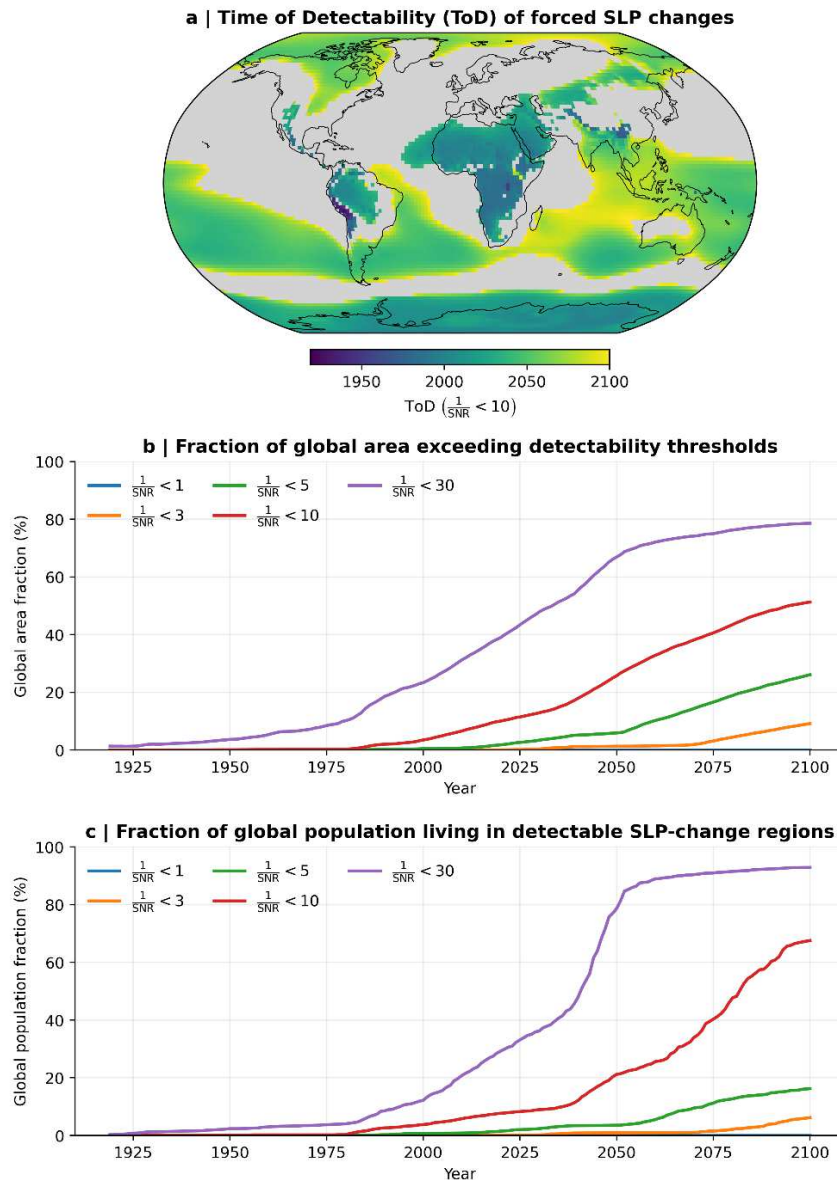


Figure 13 Regional climate emergence within the recoverability framework. a) First year at which forced sea-level pressure responses satisfy the illustrative recoverability criterion $1/\text{SNR} < 10$, estimated from expanding historical and SSP2-4.5 simulations beginning in 1850. For a single realization, this threshold corresponds to an expected agreement of $r \approx 0.30$. Grey regions do not satisfy this criterion by 2100. b) Fraction of global land area satisfying progressively more stringent recoverability criteria. Thresholds of $1/\text{SNR} < 1, 3, 5, 10$ and 30 correspond approximately to $r \approx 0.7, 0.50, 0.4, 0.30$ and 0.18 , respectively, for $N=1$. c) Fraction of the global population living in regions satisfying the same criteria.

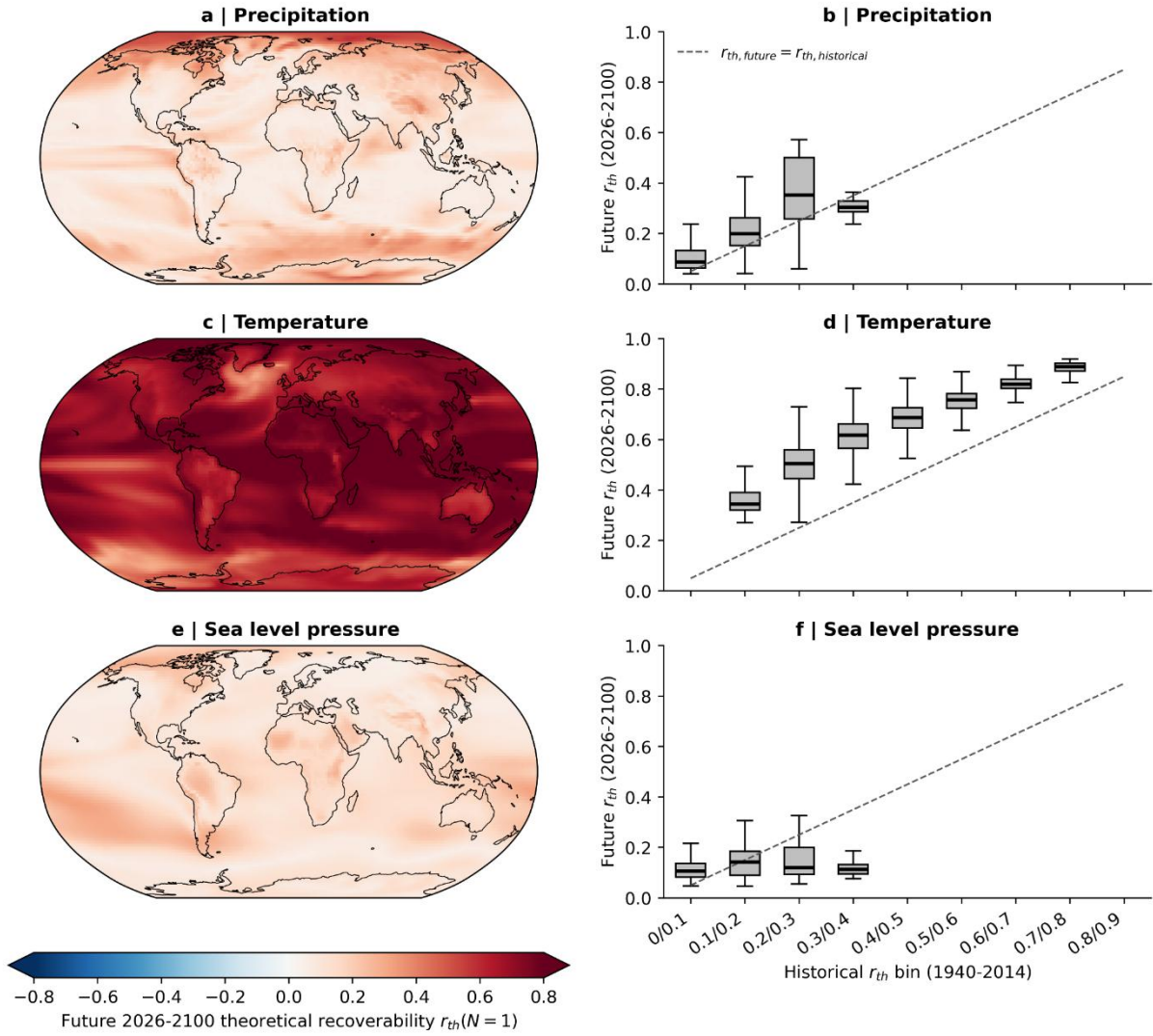


Figure 14 Projected evolution of regional climate recoverability. a,c,e, Theoretical agreement attainable from a single realization ($N=1$), $r_{th} = \frac{1}{\sqrt{1 + \frac{1}{N(SNR)}}}$, for precipitation, near-surface air temperature and sea-level pressure over 2026-2100 under SSP2-4.5. Larger values indicate regions where the externally forced response is expected to be more readily recoverable from a single climate realization. b,d,f, Distribution of future theoretical recoverability as a function of historical recoverability over 1940-2014. Grid points are grouped into historical r_{th} bins of width 0.1; boxes show the interquartile range, centre lines the median, and whiskers the 5th-95th percentile range. The one-to-one reference indicates unchanged recoverability between the historical and future periods. Values above this reference denote an increase in the agreement theoretically attainable from a single realization.

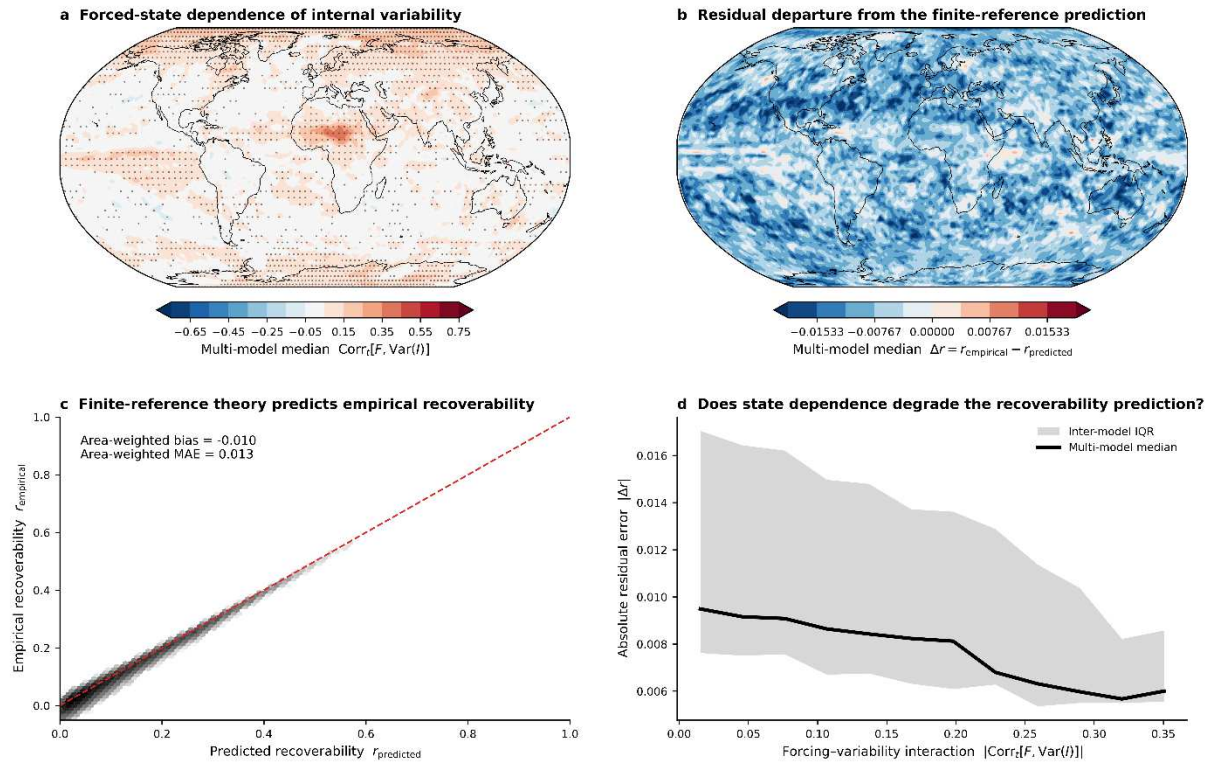


Figure 15 Robustness of the recoverability relationship to forcing-variability dependence.

a) Forced-state dependence of internal variability, quantified at each grid cell as the temporal correlation between the cross-fitted forced response, F , and the contemporaneous ensemble variance of the residual internal component, $\text{Var}(I)$. Values show the median across large-ensemble model families; stippling denotes locations where at least 75% of model families with valid estimates agree on the sign of the relationship. b) Residual departure of empirical recoverability from the prediction obtained under the independence approximation, ($\Delta r = r_{\text{empirical}} - r_{\text{predicted}}$), after accounting for uncertainty in the forced response estimated from a finite reference ensemble. c) Empirical versus predicted recoverability across model families and grid cells. The dashed red line denotes the 1:1 relationship; reported bias and mean absolute error (MAE) are area weighted. d) Absolute residual error, $|\Delta r|$, as a function of the magnitude of forcing-variability state dependence, $|\text{Corr}_t[F, \text{Var}(I)]|$. The black line show the multi-model (10 SMILES) median and shading the inter-model interquartile range. Diagnostics are calculated over 1940-2014 using repeated two-fold cross-fitting, such that the forced response and internal variability are estimated from independent subsets of ensemble members. The small residual errors and absence of a systematic increase in $|\Delta r|$ with forcing-variability state dependence indicate that the independence approximation provides a useful first-order description of annual-mean recoverability, despite detectable regional modulation of internal variability by the forced state.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.pdf](#)