

Revisiting Model – Observation Discrepancies in Recent Pacific SST Trends Using an Observational Large Ensemble

**Sieu-Cuong San^{1,2}, Clara Deser^{1*}, Pedro DiNezio^{2,3}, Adam S. Phillips¹, Tongtong Xu^{4,5},
Robert C. J. Wills⁶ and Karen A. McKinnon⁷**

October 1, 2026

¹ Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder,
CO, USA.

² The ATLAS Institute, University of Colorado Boulder, Boulder, CO, USA.

³ Departamento de Hidráulica, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad
Nacional de Córdoba, Córdoba, Argentina.

⁴ CIRES, University of Colorado Boulder, Boulder CO, USA.

⁵ NOAA Physical Sciences Division Laboratory, Boulder CO, USA.

⁶ Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstrasse 16, Zurich,
Switzerland.

⁷ Department of Atmospheric and Oceanic Sciences, Department of Statistics and Data Science,
Institute of the Environment and Sustainability, University of California, Los Angeles.

* Corresponding author: Clara Deser (cdeser@ucar.edu)

21 **Key Points:**

- 22 • Earth system models generally fail to simulate the observed pattern of Pacific sea surface
23 temperature trends in recent decades.
- 24 • A novel Observational Large Ensemble approach is used to generate plausible alternative
25 realizations of observed Pacific trends.
- 26 • Accounting for alternative observational outcomes reduces the apparent model discrepancy
27 and exposes model biases in the forced response.

28

Abstract

Earth system models have difficulty simulating the observed pattern of Pacific sea surface temperature (SST) trends in recent decades. However, the observed trend is one of many that could have occurred under the same external forcing but with a different sequence of internal variability. Here, we generate synthetic realizations of observed SST trends arising from internal variability using a linear inverse model and validate our approach against a model Large Ensemble. We combine the resulting internal trend distribution with a state-of-the-art observational estimate of the forced SST response to reassess Pacific SST trends over 1979–2020. Accounting for alternative observational outcomes reduces the apparent discrepancy between observed and simulated Pacific SST trends and exposes model biases in the forced response. Our results show that comparisons based on the single observed realization can overemphasize model–observation differences and demonstrate the value of an Observational Large Ensemble for evaluating multidecadal climate trends.

Plain Language Summary

The discrepancy between Earth System model simulations and observed Pacific sea surface temperature trends in recent decades remains a long-standing puzzle. An overlooked feature of this conundrum is the fact that the observed trend is only one of many that could have occurred as a result of internal variability superimposed upon the anthropogenically forced response. Here, we generate synthetic realizations of observed trends arising from internal variability using a skillful statistical method and combine these with a recent state-of-the-art estimate of the observed forced trend component. Accounting for alternative observational outcomes associated with internal variability substantially reduces the apparent discrepancy between observed and simulated Pacific SST trends across models. Our results show that comparisons based on the single observed

realization can overstate model–observation differences and demonstrate the value of considering alternative observational outcomes for evaluating simulated multidecadal climate trends.

1 Introduction

Tropical Pacific sea surface temperature (SST) trends in recent decades show cooling in the east and warming in the west (Latif et al., 2023; Watanabe et al., 2024; Zhao & Allen, 2019), a pattern that differs markedly from the response to anthropogenic forcing simulated by many Earth system models (Lee et al., 2022; Planton et al., 2026; Rugenstein et al., 2023; Wills et al., 2022). In particular, the observed trend in the zonal SST gradient across the equatorial Pacific lies outside or in the extreme tails of the simulated trend distributions from initial-condition Large Ensembles (LEs), highlighting a substantial model–observation discrepancy (Conde et al., 2025; Seager et al., 2022; Wills et al., 2022). Similar discrepancies occur elsewhere, including the observed cooling of the Southern Ocean, which many standard-resolution models also struggle to reproduce (Kostov et al., 2018; Rye et al., 2020; Yeager et al., 2023). These differences may reflect biases in the simulated forced response, the amplitude or characteristics of internal variability, or both (Chung et al., 2024; Jiang et al., 2024; Olonscheck et al., 2020; Seager et al., 2019; Sugiura et al., 2026; Xie et al., 2026). Although recent studies indicate that increased atmospheric and oceanic model resolution can more faithfully reproduce observed SST trends (Kang et al., 2026; San et al., 2026 (submitted)), the processes responsible for these improvements remain incompletely understood. In addition, the excessive computational cost of high-resolution simulations limits ensemble size, making it difficult to robustly distinguish forced responses from internal variability (Gutowski et al., 2020).

74 An overlooked aspect of these model–observation comparisons is limited sampling of
75 internal variability on the observational side. The instrumental record contains only one realization
76 of climate evolution, yet alternative trajectories could plausibly have occurred under a different
77 sequence of internal variability superimposed upon a common forced response, in analogy with
78 the premise of a model LE (Deser et al., 2012, 2020). The apparent model–observation discrepancy
79 in recent Pacific SST trends therefore cannot be fully assessed from the single observed realization
80 alone (Jain et al., 2023). A key question is where the actual observed trend lies within the
81 distribution of trends that could plausibly have occurred in the real world. Addressing this question
82 requires separating the forced and internal components of the observed trend, and constructing a
83 synthetic ensemble of internally-generated trends.

84 One of the first efforts to build such an “Observational Large Ensemble (OBS-LE)” that
85 preserves both the spatial and temporal characteristics of internal variability was developed by
86 McKinnon and Deser (2018, 2021). Their approach was applied to terrestrial air temperature and
87 precipitation as well as global sea level pressure, and evaluated within the framework of a model
88 LE in which the forced response and distribution of internal trends are known *a priori*. Although
89 the method performed well in the extratropics, its skill was more limited in the tropics. Here, we
90 build an OBS-LE for global SSTs using a Linear Inverse Model (LIM; Penland & Sardeshmukh,
91 1995) to generate synthetic realizations of observed internal variability, and combine these with a
92 state-of-the-art estimate of the observed forced response obtained from the Forced Component
93 Estimation Statistical Method Intercomparison Project (ForceSMIP; Wills et al., 2026). The
94 ForceSMIP estimate is based on the average of all (27) statistical and machine learning methods
95 found to be skillful in a comprehensive evaluation using model LEs for which the forced response
96 is known (see Wills et al. 2026 for details); we note that ERSSTv5 (ending in 2022) was the only

SST product considered in ForceSMIP, dictating our choice of data set and end year for this study. Our OBS-LE approach enables a more incisive evaluation of trends simulated in Earth system models by distinguishing between biases in the forced response from biases in internal variability. Furthermore, it contextualizes the single observed trend within a broader range of plausible outcomes.

Using our approach, we show substantial overlap between the observed and simulated LE distributions from 19 different CMIP5 and CMIP6 models in key metrics of the Pacific SST trend pattern, including the equatorial zonal gradient, the ratio of Indo-Pacific Warm Pool (IPWP) warming to mean tropical warming, and the spatial correlation and root-mean-square (rms) amplitude. For direct comparison to Wills et al. (2022), we focus on trends over the period 1979–2020, but find similar results for 1980–2022. We further decompose the total-trend distributions into their underlying forced and internal components. We find that the models’ range of internally generated trends is broadly comparable to that inferred from the OBS-LE, whereas their forced trends exhibit systematic biases from the observed (ForceSMIP) forced response estimate. We also find that the observed trend falls within the upper tail of the OBS-LE distribution, indicating contributions from both internal variability and external forcing. Thus, the apparent model-observation discrepancy in recent Pacific SST trends can be attributed to a combination of model bias in the forced response and chance sampling of large-amplitude internal variability in observations.

2 Materials and Methods

2.1 Linear Inverse Model

To generate alternative plausible realizations of observed internal SST variability, we use a LIM (Penland & Matrosova, 2006; Penland & Sardeshmukh, 1995). The LIM represents the evolution of the climate state using a linear dynamical operator estimated from the covariability of the training data. LIMs have been widely applied to problems of climate variability and predictability, including long-term tropical trends (Coats & Karnauskas, 2017), Pacific and global predictability (Lou et al., 2020; Newman et al., 2011; Perkins & Hakim, 2020; Shin et al., 2020), modes of Pacific and global climate variability (Capotondi et al., 2026; Di Lorenzo et al., 2023; Wu et al., 2026), marine heatwaves (Wang et al., 2023; Xu et al., 2022), and interbasin interactions (Zhao et al., 2025; Zhao & Capotondi, 2026). Details of the LIM formulation and construction of the OBS-LE are provided in Texts S1 and S2 in Supporting Information S1. Trend metrics and statistical analyses are defined in Text S3 in Supporting Information S1. We validate the skill of the LIM using the 50-member CESM2-LE (Rodgers et al., 2021) in a “perfect-model” testbed framework (see Text S4 and Figures S1–S8 in Supporting Information S1).

2.2 Data

We use monthly SST from the Extended Reconstructed SST Version 5 (ERSSTv5; Huang et al., 2017) and sea surface height (SSH) from the Ocean Reanalysis System 5 (ORAS5; Zuo et al., 2019) over the period 1960–2022 to train the LIM used for constructing the OBS-LE. Note that the LIM operates on the internal component of SST and SSH variability, which is obtained by subtracting the forced component from the data (see Text S2 in Supporting Information S1 for details). For model data, we use 17 LEs from the Multi-Model Large Ensemble Archive version 2

(MMLEAv2; Maher et al., 2025), supplemented by two CESM1-based LEs: a 10-member high-resolution ($1/4^\circ$ atmosphere, $1/10^\circ$ ocean) ensemble (CESM1-HR; Chang et al., 2026) and a 40-member low-resolution (1°) ensemble counterpart (CESM1-LR; San et al., 2026). All datasets are regridded to a common $1^\circ \times 1^\circ$ grid before analysis.

3 Results

3.1 Observational Large Ensemble of Pacific SST Trends

Observed SST trends during 1979–2020 show a striking pattern of broad cooling (warming) in the eastern (western) tropical and subtropical Pacific, reminiscent of the negative phase of Interdecadal Pacific Variability (Henley et al., 2015; Power et al., 1999). This Pacific pattern is accompanied by cooling across much of the Southern Ocean and warming in the tropical Indian and Atlantic Oceans equatorward of 40°S (Figure 1a). Internal variability and external forcing (as estimated by ForceSMIP) both contribute to the observed SST trend pattern as shown in Figure S1 in Supporting Information S1 (see Wills et al. 2026 for further discussion). However, according to the OBS-LE, this is not the only SST trend pattern that could have occurred. For example, an alternative realization could have featured enhanced warming in the eastern equatorial Pacific and along the eastern North Pacific, slight cooling in the central North Pacific, and muted cooling in the Southern Ocean (Figure 1b), whereas another could have produced a similar pattern as observed but with stronger cooling in the eastern equatorial Pacific and amplified warming in the central North Pacific (Figure 1c). Importantly, both alternative realizations are consistent with the combined influence of external (anthropogenic and natural) forcing and internal variability according to our OBS-LE methodology. Indeed, they fall within the plausible (5th–95th percentile) range of the OBS-LE based on their internal IPV trend magnitudes and are therefore not statistical outliers (see Figures S2 and S3 in Supporting Information S1). An analogous range of trend

patterns occurs in the CESM2-LE, albeit with larger amplitude due to the model's likely overestimation of internal IPV trend variability (Figures S2 and S4 in Supporting Information S1).

3.2 Comparison with CESM2

We next compare the OBS-LE and CESM2-LE trend distributions in the equatorial Pacific zonal SST gradient and the ratio of IPWP warming to mean tropical warming following the metrics defined in Wills et al. 2022 (see Text S3 in Supporting Information S1). For the gradient index (Figure 1d), the observed value lies in the extreme upper tail of the CESM2-LE distribution, at the 96th percentile, indicating that such a large positive gradient trend is rarely reproduced in the available ensemble, consistent with previous studies (Planton et al., 2026; Wills et al., 2022). The observed trend also falls within the upper portion of the OBS-LE distribution, at approximately the 82nd percentile. More importantly, the CESM2-LE and OBS-LE distributions overlap substantially, indicating that many plausible alternative real-world trend outcomes would have fallen within the model range (maroon shading in Fig. 1d). For the gradient index, we define the overlapping interval as extending from the 5th percentile of the OBS-LE distribution (-0.15°C ; dotted blue line) to the 95th percentile of the CESM2-LE distribution ($+0.48^{\circ}\text{C}$; dotted gray line). Within this range, the apparent discrepancy between what the model can simulate and what the real world could plausibly have produced would be entirely attributable to sampling fluctuation rather than model bias.

The spread of internally generated trends in the gradient index is broadly comparable between CESM2-LE and OBS-LE, although the model exhibits a somewhat larger standard deviation compared to the OBS-LE (0.34°C versus 0.27°C). In contrast, the forced component of the gradient index trend differs substantially between the CESM2-LE ($-0.14 \pm 0.048^{\circ}\text{C}$ standard error) and the OBS-LE ($+0.29^{\circ}\text{C}$). These results underscore the importance of evaluating the

forced and internal contributions separately because model biases may occur in either or both components. External forcing and internal variability contribute roughly equally (55% and 45%, respectively) to the observed trend in the gradient index ($+0.53^{\circ}\text{C}$).

The ratio of IPWP warming to mean tropical warming trends (Figure 1e) shows analogous results. The observed ERSSTv5 ratio index lies in the extreme upper tail of the CESM2-LE distribution, at the 98th percentile, but remains within the upper portion of the OBS-LE distribution, at approximately the 86th percentile. The CESM2-LE and OBS-LE distributions again overlap substantially, indicating that many plausible alternative realizations of the observed climate could have produced warming ratio trends within the range simulated by CESM2-LE. As before, we define the overlapping interval as extending from the 5th percentile of the OBS-LE distribution (0.96; dotted blue line) to the 95th percentile of the CESM2-LE distribution (1.24; dotted gray line). The spread of internally generated warming ratio trends is also broadly similar between the distributions (standard deviation of 0.11 and 0.12 for CESM2-LE and OBS-LE, respectively). In contrast, the forced warming ratio trend differs between the model and observations (0.99 ± 0.015 for CESM2-LE and 1.16 for ForceSMIP).

Together, the two metrics show a consistent result: although the observed ERSSTv5 trend lies in the extreme upper tail of the distribution sampled by the CESM2-LE ensemble, the CESM2-LE and OBS-LE distributions overlap substantially, indicating that chance internal variability could be misconstrued as model bias. While the range of internally-generated trends is broadly similar in magnitude between CESM2-LE and OBS-LE, model–observation discrepancies are evident in the forced component of both metrics. The consistent results across two measures of tropical Pacific SST trend structure highlight the importance of accounting for uncertainty

associated with internal variability -- estimated here using the OBS-LE -- when evaluating model–observation differences.

3.3 Reconciling Model–Observation Discrepancies in Pacific SST Trend Indices with the OBS-LE

Motivated by the comparison between CESM2-LE and OBS-LE, we next evaluate trends in the equatorial Pacific zonal SST gradient and the warm-pool warming ratio across the full suite of 19 model LEs (Figure 2). For the gradient index (Figure 2a), the observed ERSSTv5 trend lies at or beyond the extreme upper tail of the total-trend distributions across most model LEs. In 16 of the 19 ensembles, the observed value exceeds all individual members. In the remaining three ensembles, ERSSTv5 lies at the 96th percentile of CESM2-LE (model G), the 97th percentile of GFDL-ESM2M (model K), and the 96th percentile of MPI-GE (model R). However, substantial overlap emerges when these model distributions are compared with the OBS-LE, indicating that many alternative realizations of the observed climate could plausibly have produced gradient trends within the simulated ranges.

In contrast, the forced component shows more systematic model–observation discrepancies. The forced (ensemble-mean) gradient trend across the 19 model LEs ranges from -0.28°C in ACCESS-ESM1.5 (model A) to $+0.15^{\circ}\text{C}$ in CESM1-HR (model E), and all are weaker than the ForceSMIP estimate of $+0.29^{\circ}\text{C}$. CESM1-HR provides the closest model estimate, followed by MPI-GE ($+0.10^{\circ}\text{C}$; model R) and IPSL-CM6A ($+0.06^{\circ}\text{C}$; model N). The CESM1-HR estimate is also substantially closer to ForceSMIP than that of CESM1-LR ($+0.02^{\circ}\text{C}$; model F). The improved representation of the forced SST trend in CESM1-HR relative to CESM1-LR is discussed further in San et al. (2026).

Analogous results are found for trends in the warm-pool warming ratio (Figure 2b). The observed ERSSTv5 value lies at or beyond the upper tail of the distributions in most models, exceeding all sampled members in 16 of the 19 LEs. In the remaining three models, ERSSTv5 lies at the 80th percentile of CESM1-HR (model E) and the 98th percentile of both CESM2-LE (model G) and MIROC6 (model P). However, substantial overlap again emerges when these model distributions are compared with the OBS-LE, indicating the limitations associated with comparing models to the single observed outcome.

Like the gradient index, the forced component of the trend in the ratio index shows systematic model–observation discrepancies. The ForceSMIP estimate of the observed ratio index trend is 1.16, whereas the forced component of the ratio trends in 18 of the 19 model LEs ranges from 0.93 in ACCESS-ESM1.5 (model A) to 1.08 in IPSL-CM6A (model N) and MPI-GE (model R), all below the ForceSMIP estimate. CESM1-HR is the sole exception, with a forced ratio of 1.16 that closely matches ForceSMIP.

Together, the two regional metrics indicate that the apparent model–observation discrepancy in tropical SST trend structure is substantially reduced when the range of observationally plausible outcomes associated with internal variability is considered. At the same time, the systematic differences in the forced components point to the forced response as a more robust source of model–observation disagreement.

3.4 Reconciling Model–Observation Discrepancies in SST Trend Patterns with the OBS-LE

We next assess whether the conclusions based on the two regional indices extend to the full spatial pattern of Pacific SST trends. We examine the distribution of pattern correlations and root-mean square errors (RMSEs) between the observed (ERSSTv5) trend and the trend in each member of the OBS-LE and in each member of each model LE. For the pan-Pacific (70°N–70°S)

domain, the OBS-LE reveals substantial uncertainty in the spatial pattern of SST trends that could arise from different realizations of internal variability superimposed upon the forced response (Figures 2c and 2d). Across the full OBS-LE, pattern correlations with ERSSTv5 have a median value of 0.78 and a 5th–95th percentile range of 0.03–0.93, while RMSEs have a median value of 0.34°C and a corresponding range of 0.20–0.60°C. When the model distributions are compared with this observationally plausible range, substantial overlap emerges. For pattern correlation, 85–100% of the members from each model LE fall within the 5th–95th percentile range of the OBS-LE. For RMSE, the fraction of ensemble members lying within the OBS-LE 5th–95th percentile range varies substantially among models. Some models nearly always exceed the OBS-LE 95th percentile (e.g., CanESM2, CanESM5 and UKESM1-0-LL), while others fall entirely (CESM1-HR) or nearly entirely (MPI-CMIP6) within the plausible range of observed outcomes. Similar behavior is found for the tropical Pacific and global domains: the model distributions of pattern correlation overlap substantially with the observationally plausible range, whereas those for RMSE show more heterogeneous results across models (Figures S9 and S10 in Supporting Information S1). Additionally, the results do not depend on the domain used to train the LIM (global or pan-Pacific) nor on the exact time period used to calculate trends (1979–2020 vs. 1980–2022): see Figures S11 and S12 in Supporting Information S1.

The spatial pattern diagnostics presented above do not distinguish between forced vs. internal contributions to the observed and simulated trends. Thus, we repeat these diagnostics on the internal and forced trend components separately. For internal trends, the OBS-LE and model LE distributions overlap completely, indicating that all models simulate a realistic range of patterns and amplitudes of internal trend variability (Figure S13 in Supporting Information S1). For forced trends, pattern correlations range from 0.48–0.80 across models (the majority falling between 0.65

– 0.73) and RMSE values span from 0.22– 0.68 (red circles in Figures 2c and 2d). CESM1-HR has the highest pattern correlation and one of the lowest RMSEs. These results confirm model-dependent biases in the spatial pattern and amplitude of the forced trend, and validate models’ internal trend variability, consistent with the conclusions based on the regional indices.

A notable aspect of the observed (ERSSTv5) SST trend is the high pattern correlation (0.96) and low spatial RMS difference (0.26°C) with its forced (ForceSMIP) trend component (black diamond in column T of Figures 2c and 2d). This pattern correlation lies in the upper 5th percentile of the OBS-LE distribution while the RMS difference sits at the lower 24th percentile. Similar characteristics are found for pattern correlations and RMSEs between ERSSTv5 and the models’ forced trends, which sit in the upper and lower quartiles of the models’ distributions, respectively (black diamonds in Figures 2c and 2d). This behavior should not be interpreted as evidence of a model “signal-to-noise paradox” (Scaife & Smith, 2018), since the OBS-LE shows a similar result.

There is also a strong positive pattern correlation ($r=0.84$) between the internal and forced components of the observed trend (recall Figures S1b and S1c in the Supporting Information S1), underscoring the challenge of discriminating between forced and internal contributions on the basis of spatial structure alone. We emphasize that this resemblance may be fortuitous in the sense of being specific to the period 1979–2020, although we cannot rule out that some of the methodologies in ForceSMIP may influence this result. In support of the former, we note that 41-year trends starting in the early 1960s show almost no spatial correspondence between their forced and internal components compared to those starting in the early 1950s or late 1970s, reflecting the waxing and waning of internal multi-decadal variability such as IPV (Figure S14). Future work will address the evolving nature of the relationship between internal and forced trend patterns in

the longer observational record when ForceSMIP estimates for the first half of the 20th century become available.

In summary, the pattern diagnostics together with the regional metrics show that the particular realization of observed internal variability can account for a substantial portion of the apparent model–observation discrepancy in Pacific SST trend structure. The broad OBS-LE distributions show considerably greater overlap with the model ensembles than suggested by considering the single observed realization alone. At the same time, the systematic differences between models’ ensemble-mean (forced) pattern and the observed (ForceSMIP) forced estimate indicate that biases in the forced response remain an important source of model–observation disagreement.

4 Conclusions

We have developed an observationally-based Large Ensemble (OBS-LE) to quantify the range of multidecadal (1979–2020) Pacific SST trends that could plausibly have occurred under the same external forcing but with different realizations of internal variability. We then used this ensemble to revisit the reported discrepancy between model simulations and real-world observations across 19 different CMIP5 and CMIP6 model LEs. The OBS-LE was constructed using a LIM trained on observations of SST and SSH, combined with an independent state-of-the-art estimate of the observed forced SST response from ForceSMIP. We validated the skill of the LIM in capturing the range of internally-generated multi-decadal trends using CESM2 within a perfect-model testbed framework.

Accounting for observational uncertainty due to internal variability provides important insight into the apparent model–observation discrepancy in tropical Pacific SST trend structure. Although the single observed realization lies at or beyond the extreme upper tail of most of the 19

model LEs, the OBS-LE distribution overlaps substantially with the model LE distributions. Furthermore, the observed realization falls within the upper range of the OBS-LE distribution rather than being an extreme outlier. These conclusions are consistent across key metrics of Pacific SST trend structure, including equatorial zonal gradient, warm-pool warming ratio, and spatial pattern correlation. Agreement in spatial RMSE is more limited, with most models showing a larger RMSE than the majority of OBS-LE realizations.

The OBS-LE framework also provides a more incisive tool for model assessment by separately revealing model biases in forced versus internal SST trend components. We find that the range of internal trends in the model LEs is comparable to that of the OBS-LE, consistent with Planton et al. (2026). In contrast, systematic model-observation discrepancies occur in the forced SST trend component (based on the ForceSMIP estimate of the observed forced response). The forced component in the zonal gradient index, the warm pool warming ratio, and the spatial pattern of trends in most models differ substantially from the ForceSMIP estimates, indicating that biases in the forced response remain an important source of model–observation disagreement. CESM1-HR generally shows the closest agreement with ForceSMIP and performs substantially better than its standard (1°) resolution CESM1 counterpart. This behavior suggests that increased model resolution may reduce biases in the forced SST response (San et al., 2026), although the improvement is not uniform across all metrics.

Our results depend on the accuracy of the ForceSMIP estimate of the observed forced response. As discussed in Wills et al. (2026), methods with comparable skill yield different estimates of the observed forced trend, which directly affects the partitioning between internal variability and forced response inherent in the OBS-LE. Here, we have tried to mitigate this uncertainty by adopting the ForceSMIP skillful-method average as the most robust estimate of the

forced response. While beyond the scope of this study, investigating the sensitivity of the OBS-LE to different choices of forced response estimates, whether from ForceSMIP or through other approaches, as well as to different choices of observational SST and SSH products, is an important next step. In addition, there is scope for improving the LIM formulation used here, perhaps by incorporating seasonally-dependent covariances and temporally persistent stochastic forcing.

In conclusion, our findings demonstrate that Pacific SST trend comparisons based on the single outcome sampled in the observational record can overemphasize model–observation differences, which might in turn be misconstrued as model bias. Accounting for a range of alternative observational outcomes, each consistent with the same forced response but representing different realizations of internal variability, is therefore crucial when evaluating model performance. Separating the forced and internally generated components further helps distinguish discrepancies associated with the simulated forced response from those arising simply because models and observations sample different chance realizations of internal variability. While we have targeted recent SST trends in the Pacific sector, our OBS-LE framework could be extended to other regions, time periods and climate variables for a more informed understanding of the observational record and a broader perspective on model-observation comparisons.

Acknowledgments

We thank Drs. Matthew Newman and Antonietta Capotondi for helpful discussions regarding linear inverse modeling, and the ForceSMIP participants for their methodological contributions to the estimation of the observed forced response. We acknowledge high-performance computing support from the Derecho system (<https://doi.org/10.5065/qx9a-pg09>), provided by NSF NCAR and sponsored by the National Science Foundation. SCS, PD, and AP were supported by National

Oceanic and Atmospheric Administration grant OARX431G0060. CD was supported by NSF NCAR. RCJW was supported by the Swiss National Science Foundation (Grant PCEFP2-20337).

Open Research

Links to datasets used:

ERSSTv5: <https://psl.noaa.gov/data/gridded/data.noaa.ersst.v5.html>

ORAS5: <https://www.ecmwf.int/en/forecasts/dataset/ocean-reanalysis-system-5>

CESM2 LE: <https://www.cesm.ucar.edu/community-projects/lens2>

MMLEAv2: <https://www.cesm.ucar.edu/community-projects/mmlea/v2>

ForceSMIP: <https://zenodo.org/records/17106679>

LIM code: <https://github.com/Tongtong-Xu-PSL/LIM>

Conflict of Interest Disclosure

The authors declare no competing interests.

References

- Capotondi, A., Xu, T., Newman, M., Di Lorenzo, E., & Vimont, D. (2026). Tropical and North Pacific coupling at decadal timescales shapes basin variability and marine heatwaves. *Communications Earth & Environment*. <https://doi.org/10.1038/s43247-026-03864-7>
- Chang, P., Fu, D., Liu, X., Castruccio, F. S., Prein, A. F., Danabasoglu, G., et al. (2026). Future extreme precipitation amplified by intensified mesoscale moisture convergence. *Nature Geoscience*, 19(1), 33–41. <https://doi.org/10.1038/s41561-025-01859-1>
- Chung, E.-S., Kim, S.-J., Lee, S.-K., Ha, K.-J., Yeh, S.-W., Kim, Y. S., et al. (2024). Tropical eastern Pacific cooling trend reinforced by human activity. *Npj Climate and Atmospheric Science*, 7(1), 170. <https://doi.org/10.1038/s41612-024-00713-2>

- 390 Coats, S., & Karnauskas, K. B. (2017). Are Simulated and Observed Twentieth Century Tropical Pacific Sea
391 Surface Temperature Trends Significant Relative to Internal Variability? *Geophysical Research Letters*,
392 44(19), 9928–9937. <https://doi.org/10.1002/2017GL074622>
- 393 Conde, I., Zika, J., Holmes, R. M., & Ramesh, N. (2025). Observed Increase in Tropical Pacific Ocean Surface
394 Cold-Warm Contrast Is Well Outside Model-Simulated Range. *Geophysical Research Letters*, 52(15),
395 e2025GL116409. <https://doi.org/10.1029/2025GL116409>
- 396 Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., et al. (2020). Insights from Earth
397 system model initial-condition large ensembles and future prospects. *Nature Climate Change*, 10(4), 277–
398 286. <https://doi.org/10.1038/s41558-020-0731-2>
- 399 Deser, Clara. (2020). “Certain Uncertainty: The Role of Internal Climate Variability in Projections of Regional
400 Climate Change and Risk Management.” *Earth’s Future*, 8(12), e2020EF001854.
401 <https://doi.org/10.1029/2020EF001854>
- 402 Deser, Clara, Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: the role of
403 internal variability. *Climate Dynamics*, 38(3), 527–546. <https://doi.org/10.1007/s00382-010-0977-x>
- 404 Deser, Clara, Phillips, A. S., Simpson, I. R., Rosenbloom, N., Coleman, D., Lehner, F., et al. (2020). Isolating the
405 Evolving Contributions of Anthropogenic Aerosols and Greenhouse Gases: A New CESM1 Large
406 Ensemble Community Resource. <https://doi.org/10.1175/JCLI-D-20-0123.1>
- 407 Di Lorenzo, E., Xu, T., Zhao, Y., Newman, M., Capotondi, A., Stevenson, S., et al. (2023). Modes and Mechanisms
408 of Pacific Decadal-Scale Variability. *Annual Review of Marine Science*, 15, 249–275.
409 <https://doi.org/10.1146/annurev-marine-040422-084555>
- 410 Gutowski, W. J., Ullrich, P. A., Hall, A., Leung, L. R., O’Brien, T. A., Patricola-DiRosario, C. M., et al. (2020). The
411 Ongoing Need for High-Resolution Regional Climate Models: Process Understanding and Stakeholder
412 Information. <https://doi.org/10.1175/BAMS-D-19-0113.1>
- 413 Henley, B. J., Gergis, J., Karoly, D. J., Power, S., Kennedy, J., & Folland, C. K. (2015). A Tripole Index for the
414 Interdecadal Pacific Oscillation. *Climate Dynamics*, 45(11), 3077–3090. [https://doi.org/10.1007/s00382-](https://doi.org/10.1007/s00382-015-2525-1)
415 015-2525-1

- Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., et al. (2017). Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. <https://doi.org/10.1175/JCLI-D-16-0836.1>
- Jain, S., Scaife, A. A., Shepherd, T. G., Deser, C., Dunstone, N., Schmidt, G. A., et al. (2023). Importance of internal variability for climate model assessment. *Npj Climate and Atmospheric Science*, 6(1), 68. <https://doi.org/10.1038/s41612-023-00389-0>
- Jiang, F., Seager, R., & Cane, M. A. (2024). A climate change signal in the tropical Pacific emerges from decadal variability. *Nature Communications*, 15(1), 8291. <https://doi.org/10.1038/s41467-024-52731-6>
- Kang, S. M., Putrasahan, D. A., Brizuela, N. G., Haak, H., Kröger, J., Marotzke, J., et al. (2026). Km-scale coupled simulation and model–observation SST trend discrepancy. *Proceedings of the National Academy of Sciences*, 123(8), e2522161123. <https://doi.org/10.1073/pnas.2522161123>
- Kostov, Y., Ferreira, D., Armour, K. C., & Marshall, J. (2018). Contributions of Greenhouse Gas Forcing and the Southern Annular Mode to Historical Southern Ocean Surface Temperature Trends. *Geophysical Research Letters*, 45(2), 1086–1097. <https://doi.org/10.1002/2017GL074964>
- Latif, M., Bayr, T., Kjellsson, J., Lübbecke, J. F., Martin, T., Nnamchi, H. C., et al. (2023). Strengthening atmospheric circulation and trade winds slowed tropical Pacific surface warming. *Communications Earth & Environment*, 4(1), 249. <https://doi.org/10.1038/s43247-023-00912-4>
- Lee, S., L’Heureux, M., Wittenberg, A. T., Seager, R., O’Gorman, P. A., & Johnson, N. C. (2022). On the future zonal contrasts of equatorial Pacific climate: Perspectives from Observations, Simulations, and Theories. *Npj Climate and Atmospheric Science*, 5(1), 82. <https://doi.org/10.1038/s41612-022-00301-2>
- Lou, J., O’Kane, T. J., & Holbrook, N. J. (2020). A Linear Inverse Model of Tropical and South Pacific Seasonal Predictability. <https://doi.org/10.1175/JCLI-D-19-0548.1>
- Maher, N., Phillips, A. S., Deser, C., Wills, R. C. J., Lehner, F., Fasullo, J., et al. (2025). The updated Multi-Model Large Ensemble Archive and the Climate Variability Diagnostics Package: new tools for the study of climate variability and change. *Geoscientific Model Development*, 18(18), 6341–6365. <https://doi.org/10.5194/gmd-18-6341-2025>
- McKinnon, K. A., & Deser, C. (2018). Internal Variability and Regional Climate Trends in an Observational Large Ensemble. <https://doi.org/10.1175/JCLI-D-17-0901.1>

McKinnon, K. A., & Deser, C. (2021). The Inherent Uncertainty of Precipitation Variability, Trends, and Extremes due to Internal Variability, with Implications for Western U.S. Water Resources.

<https://doi.org/10.1175/JCLI-D-21-0251.1>

Newman, M., Alexander, M. A., & Scott, J. D. (2011). An empirical model of tropical ocean dynamics. *Climate Dynamics*, 37(9), 1823–1841. <https://doi.org/10.1007/s00382-011-1034-0>

Olonscheck, D., Rugenstein, M., & Marotzke, J. (2020). Broad Consistency Between Observed and Simulated Trends in Sea Surface Temperature Patterns. *Geophysical Research Letters*, 47(10), e2019GL086773.

<https://doi.org/10.1029/2019GL086773>

Penland, C., & Matrosova, L. (2006). Studies of El Niño and Interdecadal Variability in Tropical Sea Surface Temperatures Using a Nonnormal Filter. <https://doi.org/10.1175/JCLI3951.1>

Penland, C., & Sardeshmukh, P. D. (1995). The Optimal Growth of Tropical Sea Surface Temperature Anomalies. Retrieved from [https://journals.ametsoc.org/view/journals/clim/8/8/1520-](https://journals.ametsoc.org/view/journals/clim/8/8/1520-0442_1995_008_1999_togots_2_0_co_2.xml)

[0442_1995_008_1999_togots_2_0_co_2.xml](https://journals.ametsoc.org/view/journals/clim/8/8/1520-0442_1995_008_1999_togots_2_0_co_2.xml)

Perkins, W. A., & Hakim, G. (2020). Linear Inverse Modeling for Coupled Atmosphere-Ocean Ensemble Climate Prediction. *Journal of Advances in Modeling Earth Systems*, 12(1), e2019MS001778.

<https://doi.org/10.1029/2019MS001778>

Planton, Y. Y., Vialard, J., Fedorov, A., Lengaigne, M., McGregor, S., & Stuecker, M. F. (2026). Correcting Underestimated Internal Variability Fails to Reconcile Models With Observed Pacific SST Gradient Strengthening. *Geophysical Research Letters*, 53(12), e2025GL120001.

<https://doi.org/10.1029/2025GL120001>

Power, S., Casey, T., Folland, C., Colman, A., & Mehta, V. (1999). Inter-decadal modulation of the impact of ENSO on Australia. *Climate Dynamics*, 15(5), 319–324. <https://doi.org/10.1007/s003820050284>

Rodgers, K. B., Lee, S.-S., Rosenbloom, N., Timmermann, A., Danabasoglu, G., Deser, C., et al. (2021). Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, 12(4), 1393–1411.

<https://doi.org/10.5194/esd-12-1393-2021>

Rugenstein, M., Dhame, S., Olonscheck, D., Wills, R. J., Watanabe, M., & Seager, R. (2023). Connecting the SST Pattern Problem and the Hot Model Problem. *Geophysical Research Letters*, 50(22), e2023GL105488.

<https://doi.org/10.1029/2023GL105488>

- 472 Rye, C. D., Marshall, J., Kelley, M., Russell, G., Nazarenko, L. S., Kostov, Y., et al. (2020). Antarctic Glacial Melt
473 as a Driver of Recent Southern Ocean Climate Trends. *Geophysical Research Letters*, 47(11),
474 e2019GL086892. <https://doi.org/10.1029/2019GL086892>
- 475 San, S.-C., DiNezio, P., Deser, C., Chang, P., Castruccio, F., Rosenbloom, N., et al. (2026). High-resolution
476 simulations reconcile observed Pacific climate trends with anthropogenic forcing. *Submitted*.
- 477 Scaife, A. A., & Smith, D. (2018). A signal-to-noise paradox in climate science. *Npj Climate and Atmospheric*
478 *Science*, 1(1), 28. <https://doi.org/10.1038/s41612-018-0038-4>
- 479 Seager, R., Cane, M., Henderson, N., Lee, D.-E., Abernathy, R., & Zhang, H. (2019). Strengthening tropical Pacific
480 zonal sea surface temperature gradient consistent with rising greenhouse gases. *Nature Climate Change*,
481 9(7), 517–522. <https://doi.org/10.1038/s41558-019-0505-x>
- 482 Seager, R., Henderson, N., & Cane, M. (2022). Persistent Discrepancies between Observed and Modeled Trends in
483 the Tropical Pacific Ocean. <https://doi.org/10.1175/JCLI-D-21-0648.1>
- 484 Shin, J., Park, S., Shin, S.-I., Newman, M., & Alexander, M. A. (2020). Enhancing ENSO Prediction Skill by
485 Combining Model-Analog and Linear Inverse Models (MA-LIM). *Geophysical Research Letters*, 47(1),
486 e2019GL085914. <https://doi.org/10.1029/2019GL085914>
- 487 Sugiura, I., Smerdon, J. E., & Seager, R. (2026). Observed Trends in the Tropical Pacific Zonal Sea Surface
488 Temperature Gradient Are Statistically Significant Based on Multiple Null Hypotheses.
489 <https://doi.org/10.1175/JCLI-D-25-0460.1>
- 490 Wang, Y., Holbrook, N. J., & Kajtar, J. B. (2023). Predictability of Marine Heatwaves off Western Australia Using a
491 Linear Inverse Model. <https://doi.org/10.1175/JCLI-D-22-0692.1>
- 492 Watanabe, M., Kang, S. M., Collins, M., Hwang, Y.-T., McGregor, S., & Stuecker, M. F. (2024). Possible shift in
493 controls of the tropical Pacific surface warming pattern. *Nature*, 630(8016), 315–324.
494 <https://doi.org/10.1038/s41586-024-07452-7>
- 495 Wills, R. C. J., Dong, Y., Proistosescu, C., Armour, K. C., & Battisti, D. S. (2022). Systematic Climate Model Biases
496 in the Large-Scale Patterns of Recent Sea-Surface Temperature and Sea-Level Pressure Change.
497 *Geophysical Research Letters*, 49(17), e2022GL100011. <https://doi.org/10.1029/2022GL100011>

- Wills, R. C. J., Deser, C., McKinnon, K. A., Phillips, A., Po-Chedley, S., Sippel, S., et al. (2026). Forced Component Estimation Statistical Method Intercomparison Project (ForceSMIP). <https://doi.org/10.1175/JCLI-D-25-0326.1>
- Wu, S., Di Lorenzo, E., Zhao, Y., Newman, M., Liu, Z., Capotondi, A., et al. (2026). Decomposition of Pacific Decadal Oscillation using linear inverse models sheds light on its dominant modes and future response. *Npj Climate and Atmospheric Science*, 9(1), 43. <https://doi.org/10.1038/s41612-025-01315-2>
- Xie, S.-P., Miyamoto, A., Peng, Q., & Luongo, M. T. (2026). Major role for internal variability in tropical Pacific warming pattern over satellite era. *Proceedings of the National Academy of Sciences*, 123(36), e2615883123. <https://doi.org/10.1073/pnas.2615883123>
- Xu, T., Newman, M., Capotondi, A., Stevenson, S., Di Lorenzo, E., & Alexander, M. A. (2022). An increase in marine heatwaves without significant changes in surface ocean temperature variability. *Nature Communications*, 13(1), 7396. <https://doi.org/10.1038/s41467-022-34934-x>
- Yeager, S. G., Chang, P., Danabasoglu, G., Rosenbloom, N., Zhang, Q., Castruccio, F. S., et al. (2023). Reduced Southern Ocean warming enhances global skill and signal-to-noise in an eddy-resolving decadal prediction system. *Npj Climate and Atmospheric Science*, 6(1), 107. <https://doi.org/10.1038/s41612-023-00434-y>
- Zhao, X., & Allen, R. J. (2019). Strengthening of the Walker Circulation in recent decades and the role of natural sea surface temperature variability. *Environmental Research Communications*, 1(2), 021003. <https://doi.org/10.1088/2515-7620/ab0dab>
- Zhao, Y., & Capotondi, A. (2026). Interactions between Tropical Atlantic and Tropical Pacific in CMIP6 Models. <https://doi.org/10.1175/JCLI-D-25-0344.1>
- Zhao, Y., Song, F., Sun, D., Dong, L., & Capotondi, A. (2025). Different roles of Indian Ocean Basin and Dipole modes in tropical Pacific climate variability. *Npj Climate and Atmospheric Science*, 8(1), 163. <https://doi.org/10.1038/s41612-025-01065-1>
- Zuo, H., Balmaseda, M. A., Tietsche, S., Mogensen, K., & Mayer, M. (2019). The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment. *Ocean Science*, 15(3), 779–808. <https://doi.org/10.5194/os-15-779-2019>

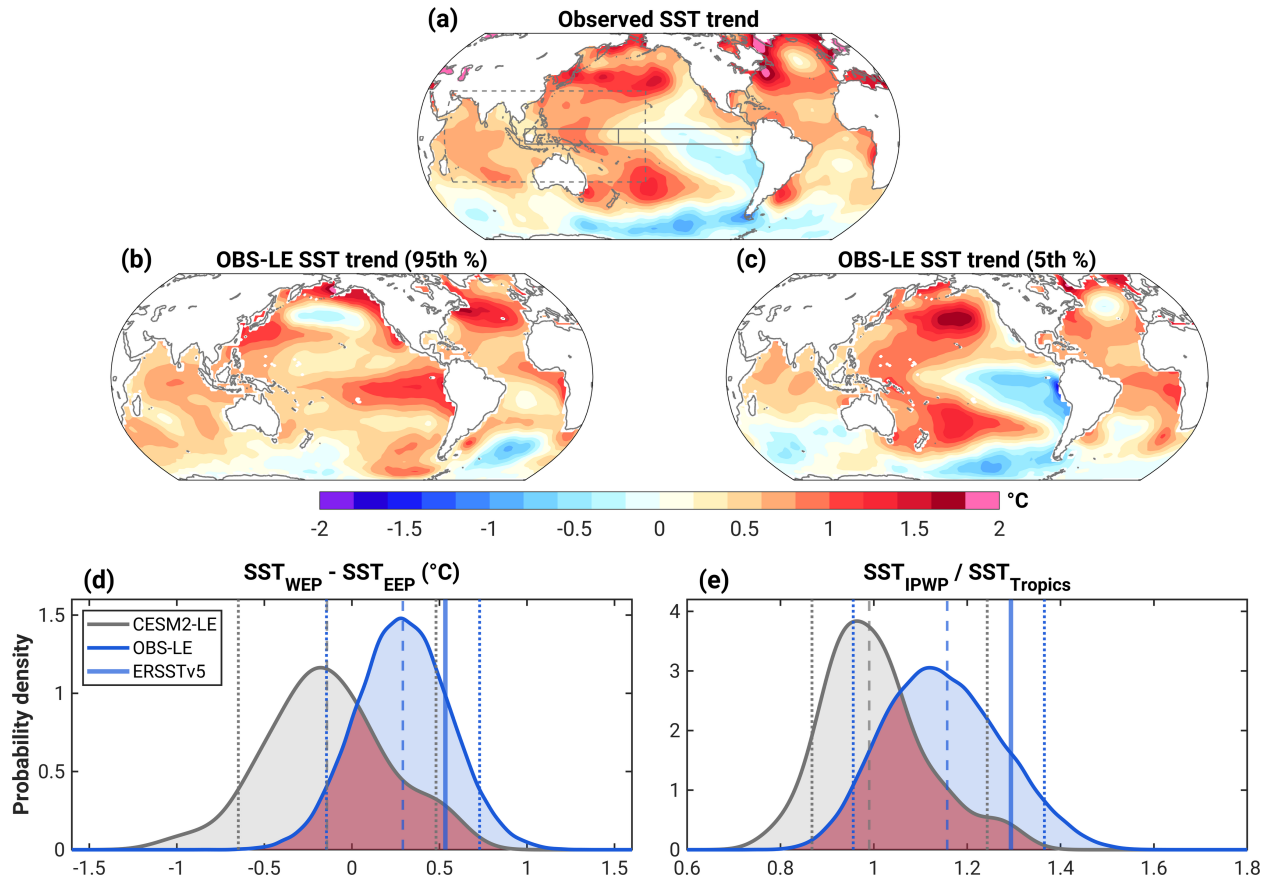


Figure 1. SST trends (°C) over 1979–2020 from (a) ERSSTv5 and (b, c) two realizations of the OBS-LE that span the plausible (5th–95th percentile) range of observed SST trends arising from internal variability superimposed upon the forced response. The realizations shown in panels b and c lie at the 95th and 5th percentiles, respectively, of the leading EOF of internal pan-Pacific SST trends in the OBS-LE, whose pattern resembles the structure of Interdecadal Pacific Variability (IPV). The solid gray boxes indicate the WEP and EEP regions used to define the east-west gradient index, and the dashed box indicates the IPWP region. Panels (d) and (e) show distributions of the equatorial Pacific zonal SST-gradient trend and the ratio of IPWP warming to tropical-mean SST warming trend, respectively, from the CESM2-LE (gray) and the OBS-LE (blue). Probability densities are estimated using Gaussian kernel density estimation (KDE) with automatically selected bandwidths. Because KDE smooths the sampled distributions, the displayed density tails may extend slightly beyond the ranges of the underlying ensemble members. Thin dotted vertical lines indicate the empirical 5th and 95th percentiles of each distribution, and dashed vertical lines indicate the corresponding forced responses (given by the ensemble-mean for CESM2-LE and ForceSMIP for the OBS-LE). The ERSSTv5 trend is indicated by the solid blue vertical line. All percentiles and percentile ranks are calculated directly from the underlying samples rather than from the fitted KDE curves.

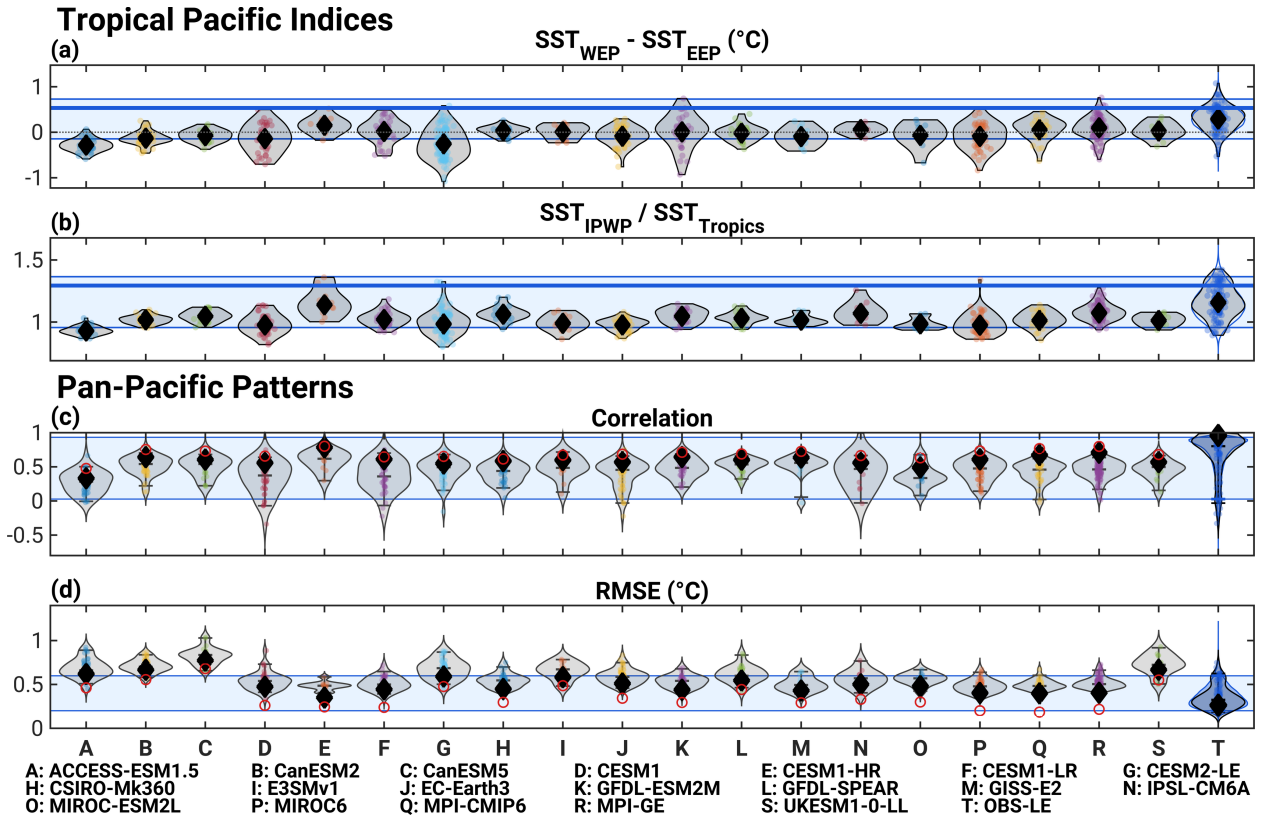


Figure 2. (Top) Trends over 1979–2020 in (a) the equatorial Pacific zonal SST gradient and (b) the ratio of IPWP warming to tropical-mean SST warming across the 19 model LEs indicated along the horizontal axis. Dots represent individual ensemble members, filled black diamonds indicate ensemble-means, and violin shapes show the KDE distributions across the members of each model LE. The OBS-LE distribution is shown at the far right, with black and blue violin outlines indicating distributions derived from the first 100 realizations and the full 50,000 member ensemble, respectively; blue dots show the first 100 realizations, and the black diamond indicates the ForceSMIP forced estimate. The thick blue horizontal line denotes the ERSSTv5 trend, and the vertical extent of the light blue shading spans the 5th–95th percentile range calculated from the full OBS-LE. (Bottom) Distributions of (c) pattern correlation and (d) RMSE between the 1979–2020 SST trends from individual members of each model LE and ERSSTv5 over the pan-Pacific domain ($70^{\circ}S$ – $70^{\circ}N$, $120^{\circ}E$ – $70^{\circ}W$; colored dots). Short horizontal black segments within each violin indicate the 5th, 50th, and 95th percentiles of the corresponding distribution. Filled black diamonds indicate the metric calculated between each model’s ensemble-mean trend map and ERSSTv5, while open red circles indicate the metric calculated between each model’s ensemble-mean trend map and ForceSMIP. The OBS-LE distribution is shown at the far right, with black and blue violin outlines indicating distributions calculated between ERSSTv5 and the first 100 and all 50,000 OBS-LE realizations, respectively; the black diamond indicates the metric calculated between ForceSMIP and ERSSTv5. The vertical extent of the light blue shading spans the 5th–95th percentile range calculated from the full OBS-LE.