



Supplement of

Combining machine learning and SMILEs to classify, better understand, and project changes in ENSO events

Nicola Maher et al.

Correspondence to: Nicola Maher (nicola.maher@colorado.edu)

The copyright of individual parts of the supplement might differ from the article licence.

S1 Datasets and additional Figures

The observational and reanalysis datasets used in this study are shown in Table S1. The single model initial-condition large ensembles (SMILEs) used in the study are found in Table S2. Figure S1 shows the same Hovmöller from Figure 2 of the main paper for SST anomalies, rather than relative SST. Figure S2 shows the relationship between the projected change of the mean-state SST gradient and the projected change in EP and CP event amplitude.

S2 Supplementary Methods

S2.1 Choice of features

In this study we tested five sets of input features (Table S3). First we use only the three niño indices 1, 2, 3 and 4 averaged over austral summer (DJF), which is the peak ENSO season. We find that the classifier can already perform reasonably well using just this limited amount of information. However, the precision and recall scores are still only 0.44/0.54 and 0.68/0.65 for CP and EP events respectively. Next we add temporal information (monthly information from October to March), which we hypothesise should help to better classify events as CP and EP events evolve differently in the temporal domain. We find that this improves the overall classifier performance, especially when considering the CP and EP events, which are now more precisely classified, with higher recall scores as well. We next add additional spatial information by splitting the niño 3 and 4 regions into halves (east and west). We perform this split at the niño 3.4 boundary. When considering the Test2 scores this slightly improves the classifier performance. Given more features may be useful when later applying the algorithm to climate models where the anomalies may not occur in exactly the same place in these models we decide to keep these additional features. We next add more temporal information as Yu and Fang (2018) suggest that including the boreal summer prior to the event may improve classification as well as another feature in the north Subtropical region as in Tseng et al. (2022). We find that the inclusion of more temporal information improves the classification, however the north Subtropical region slightly degrades it when considering the Test2 scores. Hence we choose to include the addition of extra temporal information in the final classifier, but not the north Subtropical region. We also tested a classifier that had all grid points between 160°E, 90°W, 15°S and 15°N, where all data was regridded to a 1x1 spatial grid first (full region). In this case the scores are low. We note that this result is dependent on the classifier used. We find that the nearest neighbour classifier performs poorly, which is likely due to its difficulty to perform when the data has too many dimensions. The neural network also performs poorly, likely due to too much redundant information, and over-fitting. This is an example of how 'fat data', where there are too many features, with quite a low number of events can cause an algorithm to fail. We note that a random forest algorithm, which is built to avoid over-fitting, performs well, but still not as well as our final classifier.

S2.2 Choice of algorithm

First we identified all standard algorithms used for supervised learning in python (Table S4) and tested their performance. In this case we train the classifier on all datasets bar HadISST, which we hold aside for testing. We additionally tested ensemble classifiers that used a combination of the standard algorithms (Table S4). We exclude algorithms 3, 6, and 7 due to their poor precision or recall scores for either CP or EP events. We exclude (8) QDA as it is not well suited for classification when the classes are unbalanced. We choose to keep (9) random forest over (4) decision tree as a random forest can better generalise over data than a decision tree. The final classifier used in this study is an ensemble of algorithms 1, 5 & 9 with soft voting. This algorithm performs best when considering all scores, although we note that many of the other classifiers also perform well for our dataset.

Within the final classifier we optimised the three input algorithms to find the best possible parameters for performance. As a last step we test whether a two-step classifier would perform better than a one-step classifier. The two steps for the two-step classifier are as follows:

1. Train the algorithm to classify into three categories LN, NE and EN (all El Niño events)
2. Then re classify EN into EL and CP

Theoretically this could help the classifier perform due to the low numbers of EL and CP events. However, we find that this does not improve our classifier (Table S4) and as such choose to use the simpler one-step classifier for the rest of this study.

45 S3 Shifted niño regions

Given climate models have known ENSO biases, particularly in the location of SST anomalies along the equator, we additionally classify by shifting the longitudes of the niño regions. This shift is defined as the difference in location between the maximum variability between 5N and 5S in the Pacific Ocean in the observations and the maximum variability in each individual SMILE (Table S5). We find that this does not significantly change the results in the main text except for CSIRO frequency where EP El Niños and La Niñas are now more realistically represented. The spatial patterns for each model (Figure S3) and evolution of SST anomalies (Figure S4) are very similar when applying this shift. This method additionally does not change the results for ENSO frequency (Figure S5) or amplitude (Figure S6).

S4 Extreme El Niños

S4.1 Method

55 We additionally investigate Extreme El Niño events, by including the strongest events in the observational period as their own class *strong El Niños (ST)*. The years defined as ST El Niños are 1957,1965, 1972, 1982, 1987,1997 and 2015. We choose only to include strong EP events in this category. We use the same ensemble classifier algorithm to classify these events. The algorithm performs well when the original training and evaluation are used (Table S6). However, this algorithm performs less well when we test the sensitivity to this construction. In this case the precision and recall of is significantly reduced as compared to the original classifier used in the main text (Table 3 & S6), likely due to the reduction of the number of events in the EP class. This means that the classifier is less well constrained when including ST events. Because of this we used the better constrained algorithm that does not include ST events to present the main finding of this study and discuss this additional algorithm for ST here in the Supplementary.

60 We then apply the classifier that includes extreme El Niños (ST) to the same set of SMILEs and compare the results for the EP and ST classes from this new classification. We find that the evolution of SST anomalies on the equator is similar for EP and ST events, however the ST events have much larger SST anomalies demonstrating that this classifier is now splitting the original set of EP events into weaker and stronger subsets (Figure S7).

S4.2 Results

70 When considering projections of amplitude and frequency (Figure S8) we find that the CESM-LE and CanESM5 amplitude increases occur for only the EP events, but that the CanESM2 and GFDL-ESM2M decreases occur for both types of El Niño. For frequency we find that the projected changes are confined to the ST events for CESM-LE and GFDL-ESM2M, but occur only for the EP events for CanESM5. Last, the SST and precipitation projected changes are similar for EP and ST events (Figure S9), with the changes consistently stronger for the extreme El Niños. These results are in conflict with previous work that finds an increase extreme EP events in future (Cai et al., 2014, 2018, 2021). However, based on limitations of our classification due to the small number of ST events available to train the classifier as well as clear model differences found, this warrants further investigation in future work.

References

- Balmaseda, M. A., Mogensen, K., and Weaver, A. T.: Evaluation of the ECMWF ocean reanalysis system ORAS4, *Quarterly Journal of the Royal Meteorological Society*, 139, 1132–1161, <https://doi.org/10.1002/qj.2063>, 2012.
- 80 Behringer, D. and Xue, Y.: Evaluation of the global ocean data assimilation system at NCEP: The Pacific Ocean. Eighth Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface, AMS 84th Annual Meeting, Washington State Convention and Trade Center, Seattle, Washington, 11-15, 2004.
- Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D’Andrea, F., Davini, P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.-L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.-A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.-Y., Guenet, B., Guez, Lionel, E., Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luysaert, S., Madec, G., Madeleine, J.-B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N.: Presentation and Evaluation of the IPSL-CM6A-LR Climate Model, *Journal of Advances in Modeling Earth Systems*, 12, e2019MS002010, <https://doi.org/https://doi.org/10.1029/2019MS002010>, e2019MS002010 10.1029/2019MS002010, 2020.
- 85 Cai, W., Borlace, S., Lengaigne, M., van Rensch, P., Collins, M., Vecchi, G., Timmermann, A., Santoso, A., McPhaden, M. J., Wu, L., England, M. H., Wang, G., Guilyardi, E., and Jin, F.-F.: Increasing frequency of extreme El Niño events due to greenhouse warming, *Nature Climate Change*, 4, 111–116, <https://doi.org/10.1038/nclimate2100>, 2014.
- 95 Cai, W., Wang, G., Dewitte, B., Wu, L., Santoso, A., Takahashi, K., Yang, Y., Carréric, A., and McPhaden, M. J.: Increased variability of eastern Pacific El Niño under greenhouse warming, *Nature*, 564, 201–206, <https://doi.org/10.1038/s41586-018-0776-9>, 2018.
- Cai, W., Santoso, A., Collins, M., Dewitte, B., Karamperidou, C., Kug, J.-S., Lengaigne, M., McPhaden, M. J., Stuecker, M. F., Taschetto, A. S., Timmermann, A., Wu, L., Yeh, S.-W., Wang, G., Ng, B., Jia, F., Yang, Y., Ying, J., Zheng, X.-T., Bayr, T., Brown, J. R., Capotondi, A., Cobb, K. M., Gan, B., Geng, T., Ham, Y.-G., Jin, F.-F., Jo, H.-S., Li, X., Lin, X., McGregor, S., Park, J.-H., Stein, K., Yang, K., Zhang, L., and Zhong, W.: Changing El Niño-Southern Oscillation in a warming climate, *Nature Reviews Earth Environment*, <https://doi.org/10.1038/s43017-021-00199-z>, 2021.
- 100 Carton, J. A., Chepurin, G. A., and Chen, L.: SODA3: A New Ocean Climate Reanalysis, *Journal of Climate*, 31, 6967 – 6983, <https://doi.org/10.1175/JCLI-D-18-0149.1>, 2018.
- 105 Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth system model initial-condition large ensembles and future prospects, *Nature Climate Change*, 10, 277–286, <https://doi.org/10.1038/s41558-020-0731-2>, 2020.
- Hirahara, S., Ishii, M., and Fukuda, Y.: Centennial-scale sea surface temperature analysis and its uncertainty, *Journal of Climate*, 27, 57–75, 2014.
- 110 Huang, B., Banzon, V. F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T. C., Smith, T. M., Thorne, P. W., Woodruff, S. D., and Zhang, H.-M.: Extended Reconstructed Sea Surface Temperature (ERSST), Version 4., NOAA National Centers for Environmental Information, <https://doi.org/doi:10.7289/V5KD1VVF>, 2015.
- Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M.: NOAA Extended Reconstructed Sea Surface Temperature (ERSST), Version 5, NOAA National Centers for Environmental Information, <https://doi.org/doi:10.7289/V5T72FNM>, 2017.
- 115 Ishii, M. A., Sugimoto, S. S., and Matsumoto, T.: Objective Analyses of Sea-Surface Temperature and Marine Meteorological Variables for the 20th Century using ICOADS and the Kobe Collection, *International Journal of Climatology*, 2005.
- Jeffrey, S. et al.: Australia’s CMIP5 submission using the CSIRO-Mk3.6 model, *Australian Meteorological and Oceanographic Journal*, 63, 1–13, 2012.
- 120 Kaplan, A., Cane, M., Kushnir, Y., Clement, A., Blumenthal, M., and Rajagopalan, B.: Analyses of global sea surface temperature 1856-1991, *Journal of Geophysical Research*, 103, 18 567–18 589, 1998.
- 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 American Meteorological Society*, 96, 1333–1349, <https://doi.org/10.1175/BAMS-D-13-00255.1>, 2015.
- 125 Kirchmeier-Young, M., Zwiers, F., and Gillett, N.: Attribution of Extreme Events in Arctic Sea Ice Extent, *Journal of Climate*, 30, 553–571, <https://doi.org/10.1175/JCLI-D-16-0412.1>, 2017.

- Kushner, P. J., Mudryk, L. R., Merryfield, W., Ambadan, J. T., Berg, A., Bichet, A., Brown, R., Derksen, C., Déry, S. J., Dirkson, A., Flato, G., Fletcher, C. G., Fyfe, J. C., Gillett, N., Haas, C., Howell, S., Laliberté, F., McCusker, K., Sigmond, M., Sospedra-Alfonso, R., Tandon, N. F., Thackeray, C., Tremblay, B., and Zwiers, F. W.: Canadian snow and sea ice: assessment of snow, sea ice, and related climate processes in Canada's Earth system model and climate-prediction system, *The Cryosphere*, 12, 1137–1156, <https://doi.org/10.5194/tc-12-1137-2018>, 2018.
- Köhl, A.: Evaluation of the GECCO2 Ocean Synthesis: Transports of Volume, Heat and Freshwater in the Atlantic, *Quarterly Journal of the Royal Meteorological Society*, 141, 166–181, <https://doi.org/doi:10.1002/qj.2347>, 2015.
- Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., Kröger, J., Takano, Y., Ghosh, R., Hedemann, C., Li, C., Li, H., Manzini, E., Notz, D., Putrasahan, D., Boysen, L., Claussen, M., Ilyina, T., Olonscheck, D., Raddatz, T., Stevens, B., and Marotzke, J.: The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability, *Journal of Advances in Modeling Earth Systems*, 11, 2050–2069, <https://doi.org/10.1029/2019MS001639>, 2019.
- Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *Journal of Geophysical Research*, 108, <https://doi.org/doi:10.1029/2002JD002670>, 2003.
- Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., and Schlax, M. G.: Daily High-Resolution-Blended Analyses for Sea Surface Temperature, *Journal of Climate*, 20, 5473–5496, 2007.
- Rodgers, K. B., Lin, J., and Frölicher, T. L.: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model, *Biogeosciences*, 12, 3301–3320, 2015.
- Smith, T. M., Reynolds, R. W., Peterson, T. C., and Lawrimore, J.: Extended Reconstructed Sea Surface Temperature (ERSST) Monthly Analysis, Version 3b, NOAA National Climatic Data Center., <https://doi.org/doi:10.7289/V5Z31WJ4>, 2010.
- Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., von Salzen, K., Yang, D., and Winter, B.: The Canadian Earth System Model version 5 (CanESM5.0.3), *Geoscientific Model Development*, 12, 4823–4873, <https://doi.org/10.5194/gmd-12-4823-2019>, 2019.
- Systems, R. S.: AMSRE 25km gridded SST data set. Ver. 7.0. PO.DAAC, CA, USA, <https://doi.org/https://doi.org/10.5067/GHAMS-2GR07>, 2014.
- Tseng, Y.-h., Huang, J.-H., and Chen, H.-C.: Improving the Predictability of Two Types of ENSO by the Characteristics of Extratropical Precursors, *Geophysical Research Letters*, 49, e2021GL097190, <https://doi.org/https://doi.org/10.1029/2021GL097190>, 2022.
- Wang, B., Luo, X., Yang, Y.-M., Sun, W., Cane, M. A., Cai, W., Yeh, S.-W., and Liu, J.: Historical change of El Niño properties sheds light on future changes of extreme El Niño, *Proceedings of the National Academy of Sciences*, 116, 22512–22517, <https://doi.org/10.1073/pnas.1911130116>, 2019.
- Yu, J.-Y. and Fang, S.-W.: The Distinct Contributions of the Seasonal Footprinting and Charged-Discharged Mechanisms to ENSO Complexity, *Geophysical Research Letters*, 45, 6611–6618, <https://doi.org/https://doi.org/10.1029/2018GL077664>, 2018.
- Zuo, H., Balmaseda, M. A., de Boisseson, E., Hirahara, S., Chrust, M., and De Rosnay, P.: generic ensemble generation scheme for data assimilation and ocean analysis, ECMWF Tech Memo, <https://doi.org/10.21957/cub7mq0i4>, 2017.
- Zuo, H., Balmaseda, M. A., Tietsche, S., Mogensen, K., and Mayer, M.: The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment, *Ocean Sciences*, 15, 779–808, <https://doi.org/10.5194/os-15-779-2019>, 2019.

Table S1. Input observational datasets used in this study.

dataset	years	reference
AMSREv07	2002-2010	Systems (2014)
COBEv1	1896-2019	Ishii et al. (2005)
COBESST2	1896-2018	Hirahara et al. (2014)
HadISST	1896-2018	Rayner et al. (2003)
ERSSTv3b	1896-2018	Smith et al. (2010)
ERSSTv4	1896-2018	Huang et al. (2015)
ERSSTv5	1896-2019	Huang et al. (2017)
GECCO2	1948-2016	Köhl (2015)
GODAS	1980-2019	Behringer and Xue (2004)
kaplan	1896-2019	Kaplan et al. (1998)
OISST	1982-2018	Reynolds et al. (2007)
ORAS4	1958-2016	Balmaseda et al. (2012)
ORAs5 (5 ensemble members)	1979-2017	Zuo et al. (2019, 2017)
soda3.11.2	1980-2015	Carton et al. (2018)
soda3.12.2	1980-2017	Carton et al. (2018)
soda3.4.2	1980-2018	Carton et al. (2018)
soda3.6.1	1980-2008	Carton et al. (2018)
soda3.7.2	1980-2016	Carton et al. (2018)

Table S2. SMILEs used in this study, the length of their historical period, the forcing scenario used, and the reference for the dataset. We note that the CMIP5 SMILEs are included in the Multi Model Large Ensemble Archive (MMLEA Deser et al., 2020)

SMILE	historical period	scenario	ensemble size	reference
MPI-GE	1850-2005	RCP8.5	100	Maher et al. (2019)
CESM-LE	1920-2005	RCP8.5	40	Kay et al. (2015)
CanESM2	1950-2020	RCP8.5	50	Kirchmeier-Young et al. (2017); Kushner et al. (2018)
GFDL-ESM2M	1950-2005	RCP8.5	30	Rodgers et al. (2015)
CSIRO	1850-2005	RCP8.5	30	Jeffrey et al. (2012)
CanESM5	1850-2014	SSP370	23	Swart et al. (2019)
IPSL-CM6A	1850-2014	n/a	26	Boucher et al. (2020)

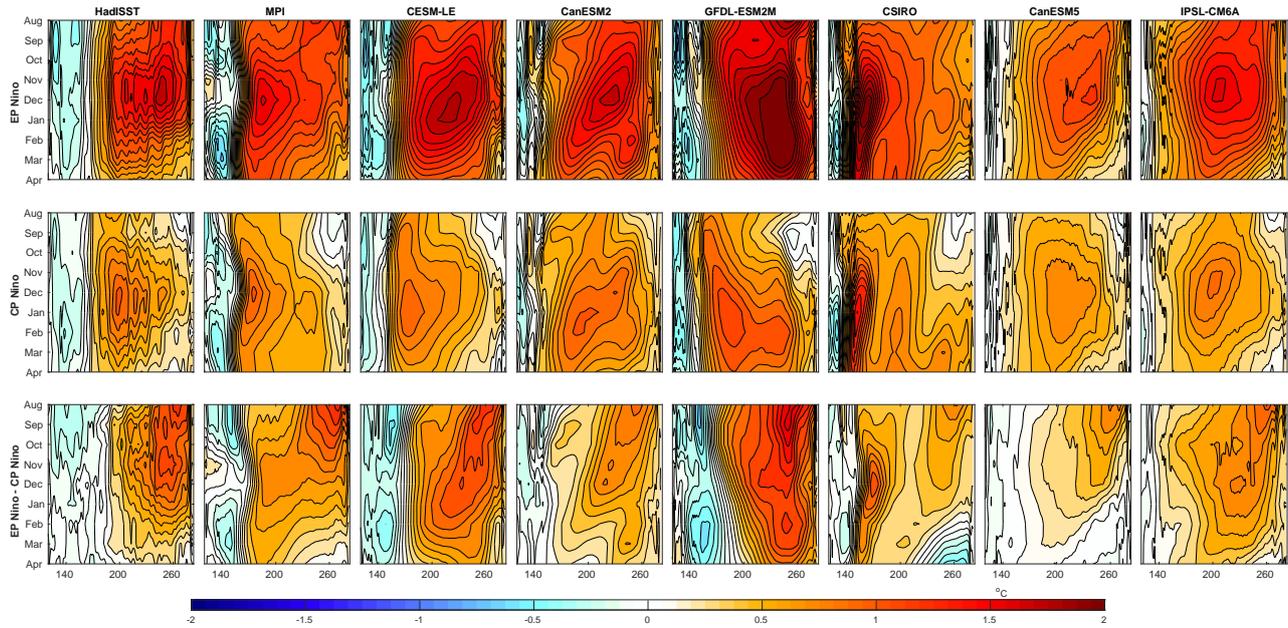


Figure S1. Hovmöller of SST along the equator in the Pacific Ocean for composites of EP and CP El Niños, and EP minus CP El Niños (top, middle and bottom row respectively). Shown for HadISST observations (left column) and each individual SMILE (in order of appearance; MPI-GE, CESM-LE, CanESM2, GFDL-ESM2M, CSIRO, CanESM5 and IPSL-CM6A). SST is averaged between 5N and 5S and shown for August to April. SMILE data has the forced response (ensemble mean) removed prior to calculation, HadISST is detrended using a second order polynomial then each months average is removed. The time period used is all of the historical, which is shown for the observations in Table S1 and SMILEs in Table S2.

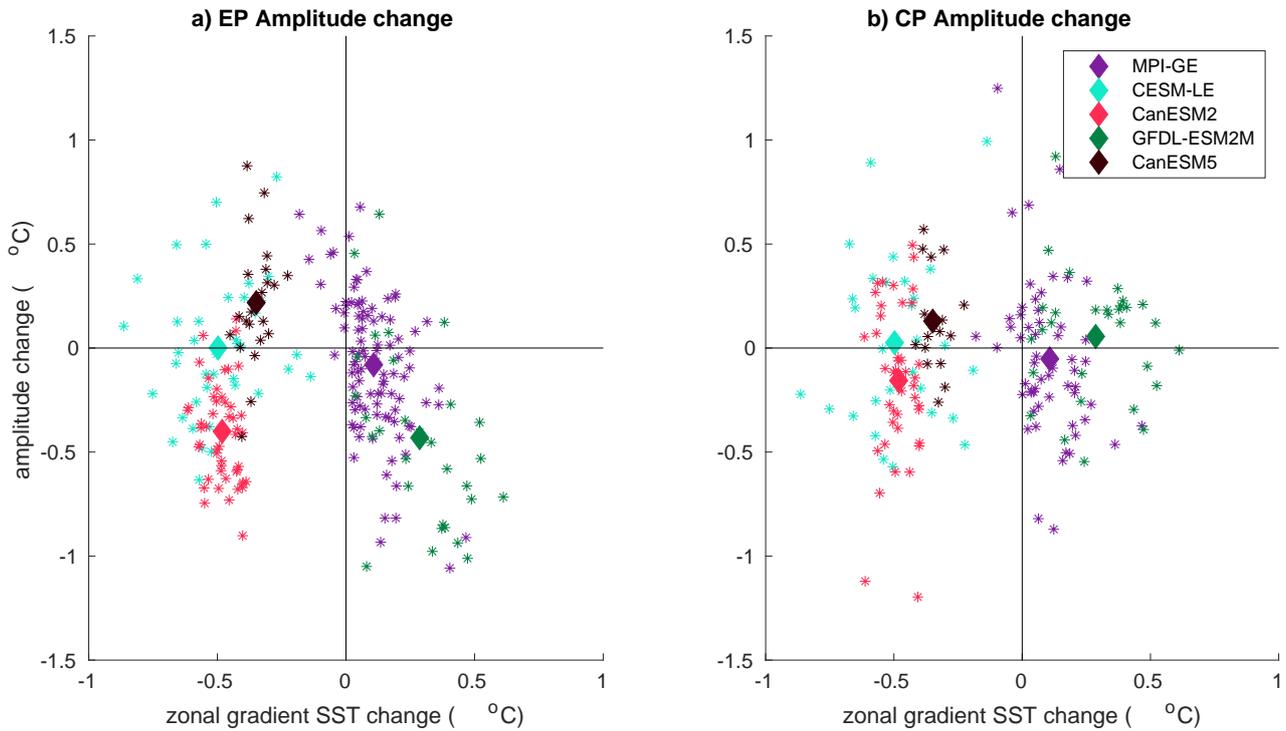


Figure S2. Relationship between future projected change in ENSO amplitude and the projected change of the mean-zonal gradient. Shown for a) EP El Niños and b) CP El Niños. Amplitude is calculated as the November, December, January mean for the region 160E to 80W between 5N and 5S after the ensemble mean has been removed for each event. All changes are calculated over the period 2069-2098 compared to 1950-1979. The zonal mean gradient is calculated as the difference in SST in the western Pacific (5S-5N and 150-180E) minus the eastern Pacific (5S-5N and 80-180W) as in Wang et al. (2019).

Table S3. Scores for different feature inputs tested. Scores and test types are defined in section 2.2 of the main paper. The full tropical Pacific is 160E-80W and 15S-15N. The North Subtropical Pacific (NSTrop) is defined as 175-120W and 5-30N. O-M denotes the months October to March. J-M denotes the months June to March. N denotes the niño regions. Bold font describes the final classifier used.

Test type	Features used	Accuracy	CVS	P-CP	P-EP	P-LN	P-NE	R-CP	R-EP	R-LN	R-NE
Test 1	DJF-mean N12,3,4	0.74	0.73	0.44	0.68	1	0.77	0.54	0.65	0.54	0.89
	O-M N12,3,4	0.93	0.91	0.92	1	1	0.9	0.92	1	0.77	0.98
	O-M N12.3E/W,4E/W	0.92	0.91	0.92	1	1	0.87	0.92	1	0.93	0.98
	J-M N12.3E/W,4E/W	0.96	0.96	1	1	1	0.93	0.92	1	0.85	1
	J-M N12.3E/W,4E/W, NSTrop	0.95	0.96	1	1	1	0.91	0.92	1	0.81	1
Test 2 mean (min/max)	DJF-mean N12,3,4	0.58 (0.54/0.62)	0.62 (0.60/0.65)	0.25 (0.09/0.41)	0.53 (0.45/0.62)	0.64 (0.42/0.81)	0.65 (0.54/0.8)	0.2 (0.11/0.29)	0.58 (0.27/0.76)	0.65 (0.52/0.76)	0.67 (0.54/0.71)
	O-M N12,3,4	0.73 (0.68/0.79)	0.94 (0.93/0.95)	0.44 (0.26/0.6)	0.66 (0.5/0.92)	0.83 (0.61/0.98)	0.77 (0.65/0.89)	0.34 (0.14/0.64)	0.71 (0.57/0.86)	0.79 (0.67/0.98)	0.82 (0.73/0.96)
	O-M N12.3E/W,4E/W	0.74 (0.69/0.80)	0.94 (0.93/0.95)	0.50 (0.29/0.73)	0.68 (0.58/0.97)	0.82 (0.52/1)	0.77 (0.66/0.87)	0.34 (0.14/0.79)	0.77 (0.63/0.89)	0.80 (0.68/0.98)	0.81 (0.69/0.96)
	J-M N12.3E/W,4E/W	0.77 (0.68/0.87)	0.97 (0.96/0.98)	0.75 (0.54/0.95)	0.73 (0.60/0.93)	0.80 (0.51/1)	0.77 (0.63/0.87)	0.60 (0.46/0.86)	0.80 (0.6/1)	0.71 (0.60/0.94)	0.83 (0.7/0.99)
	J-M N12.3E/W,4E/W, NSTrop	0.76 (0.66/0.86)	0.98 (0.97/0.98)	0.74 (0.5/1)	0.70 (0.58/0.90)	0.82 (0.53/1)	0.78 (0.62/0.88)	0.58 (0.43/0.81)	0.77 (0.6/1)	0.72 (0.60/0.98)	0.83 (0.65/0.98)
	O-M Full tropical Pacific	0.44	0.66	0.32	0.38	0.35	0.66	0.54	1	0.23	0.33
	Nearest Neighbours n=1	0.35	0.61	0.21	0.29	0.3	0.54	0.54	0.45	0.23	0.33
	Neural Network	0.16	0.41	0	0.16	0	0	0	1	0	0
	Random Forest	0.87	0.89	0.77	0.95	0.95	0.84	0.77	0.95	0.73	0.92

Table S4. Scores for different algorithms tested. Scores are defined in section 2.2 of the main text.

Algorithm	Accuracy	CVS	P-CP	P-EP	P-LN	P-NE	R-CP	R-EP	R-LN	R-NE
(1) NearestNeighbours	0.92	0.93	0.92	1	0.95	0.89	0.92	0.95	0.73	0.98
(2) LinearSVM	0.85	0.86	0.8	0.95	1	0.8	0.62	0.9	0.58	0.98
(3) RBFSVM	0.54	0.53	1	0	0	0.53	0.15	0	0	1
(4) DecisionTree	0.83	0.8	0.8	0.94	1	0.78	0.62	0.8	0.62	0.97
(5) NeuralNet	0.94	0.92	0.92	1	1	0.91	0.92	1	0.81	0.98
(6) AdaBoost	0.52	0.55	0.25	0	0.78	0.67	0.77	0	0.27	0.73
(7) NaiveBayes	0.77	0.74	0.5	0.75	0.83	0.86	0.85	0.75	0.77	0.77
(8) QDA	0.84	0.82	1	1	1	0.76	0.85	0.55	0.65	1
(9) RandomForest	0.79	0.8	0.78	0.73	1	0.77	0.53	0.8	0.57	0.92
Hard Vote (1,5,9)	0.95	0.96	1	1	1	0.91	0.92	1	0.81	1
Hard Vote (1,2,5,9)	0.95	0.94	1	1	1	0.91	0.92	1	0.81	1
Hard Vote (1,2,3,5,7,8,9)	0.91	0.94	1	1	1	0.85	0.85	0.95	0.69	1
2-step soft vote (1,5,9)	0.95	n/a	1	0.91	1	0.94	0.92	1	0.85	0.98
FINAL Soft vote (1,5,9)	0.96	0.95	1	1	1	0.93	0.92	1	0.85	1

Table S5. Frequency of events (as a percentage) in the historical period for observations (HadISST) and the SMILEs as well as the correlation between EP and CP patterns shown for the shifted center of variability. The mean frequency and correlation across each ensemble is shown with the minimum and maximum in brackets. The time period used is all of the historical, which is shown for the observations in Table S1 and SMILEs in Table S2.

Model	EP no ev shift	CP no ev shift	LN no ev shift	EP/CP shift-corr	shift longitude
HadISST	16.1	11.2	21.0	0.85	na
MPI-GE	15.7 (7.1/20)	5.8 (0.6/12.9)	17.5 (7.1/23.9)	0.75 (0.29/0.89)	-11
CESM-LE	22.4 (15.3/31.8)	6.1 (2.4/12.9)	23.4 (17.6/35.3)	0.75 (0.45/0.89)	-8
CanESM2	19.1 (12.9/27.1)	7.3 (0/15.7)	21.9 (14.3/28.6)	0.75(0.35/0.88)	-11
GFDL-ESM2M	20.3 (12.7/27.3)	14.2 (7.3/21.8)	26.5 (14.5/36.4)	0.68 (0.53/0.84)	-12
CSIRO	14.9 (10.3/21.3)	6.6 (1.9/11.6)	17.0 (10.3/21.9)	0.78 (0.54/0.90)	-26
CanESM5	9.0 (4.3/12.8)	4.7 (1.8/7.9)	15.8 (11.0/22.0)	0.78 (0.63/0.89)	-10
IPSL-CM6A	18.7 (14.6/22.6)	3.4 (0.6/7.3)	20 (15.2/25)	0.67 (0.16/0.86)	-11

Table S6. Minimum, mean and maximum scores for the ensemble classifier. Test 1 uses all available data, with HadISST kept aside for testing. Test 2 uses the longer datasets, ERSST, COBE, Kaplan and HadISST for training and testing. The data is split so that the augmented events must all occur in the same section of the data. To complete this we use the python function *train test split*. 10 splits are manually chosen to ensure that they sample events from across the time-dimension and have a reasonable amount of each type of event.

Test	Min/Max score	Accuracy	clf	P-CP	P-EP	P-LN	P-NE	P-ST	R-CP	R-EP	R-LN	R-NE	R-ST
Test 1		0.96	0.95	1	1	1	0.93	1	0.92	1	0.85	1	1
Test 2 w/check	Min	0.65	0.96	0.5	0.26	0.46	0.67	0.5	0.43	0.29	0.43	0.74	0.21
	Mean	0.72	0.97	0.79	0.4	0.79	0.76	0.77	0.66	0.56	0.69	0.82	0.46
	Max	0.82	0.98	0.95	0.61	0.97	0.85	1	0.95	0.89	0.96	0.96	0.62

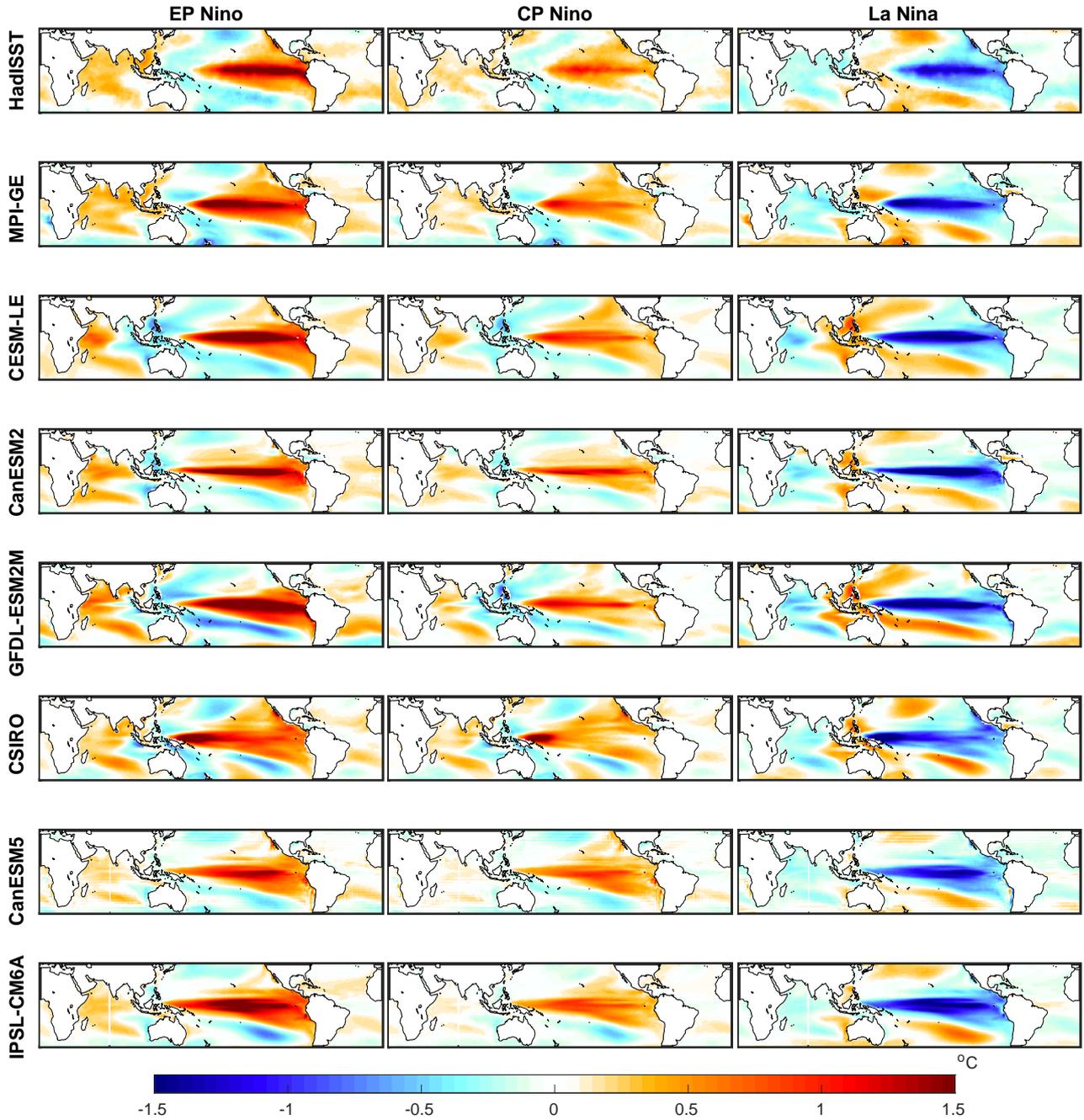


Figure S3. SST pattern for composites of EP, CP, and LN events (left, middle and right columns respectively) for the shifted centers of variability. Shown for HadISST observations (top row) and each individual SMILE (in order of appearance; MPI-GE, CESM-LE, CanESM2, GFDL-ESM2M, CSIRO, CanESM5 and IPSL-CM6A). SST pattern is shown for the November, December, January average. SMILE data has the forced response (ensemble mean) removed prior to calculation, HadISST is detrended using a second order polynomial then each months average is removed. The time period used is all of the historical, which is shown for the observations in Table S1 and SMILEs in Table S4.

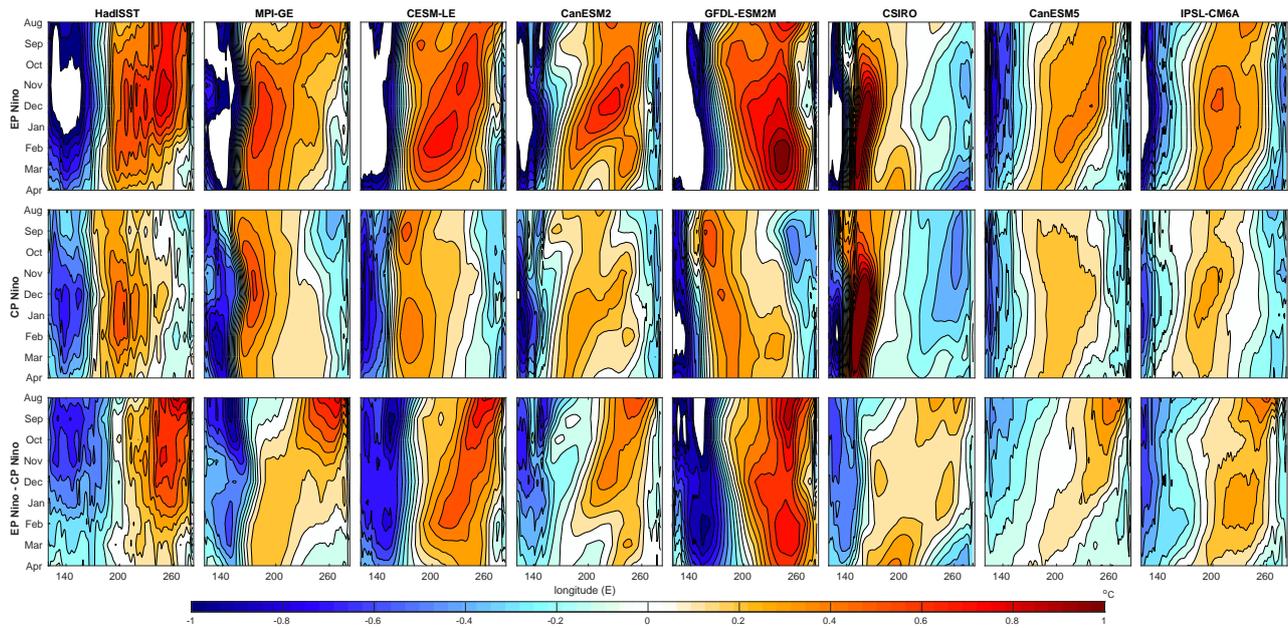


Figure S4. Hovmöller of relative SST along the equator in the Pacific Ocean for composites of EP, CP El Niños, and EP minus CP El Niños (top, middle and bottom row respectively) for the shifted centers of variability. Shown for HadISST observations (left column) and each individual SMILE (in order of appearance; MPI-GE, CESM-LE, CanESM2, GFDL-ESM2M, CSIRO, CanESM5 and IPSL-CM6A). SST is averaged between 5N and 5S and shown for August to April. SMILE data has the forced response (ensemble mean) removed prior to calculation, HadISST is detrended using a second order polynomial then each months average is removed. The time period used is all of the historical, which is shown for the observations in Table S1 and SMILEs in Table S4. Relative SST is calculated by removing the average SST between 120E and 280E individually for each month.

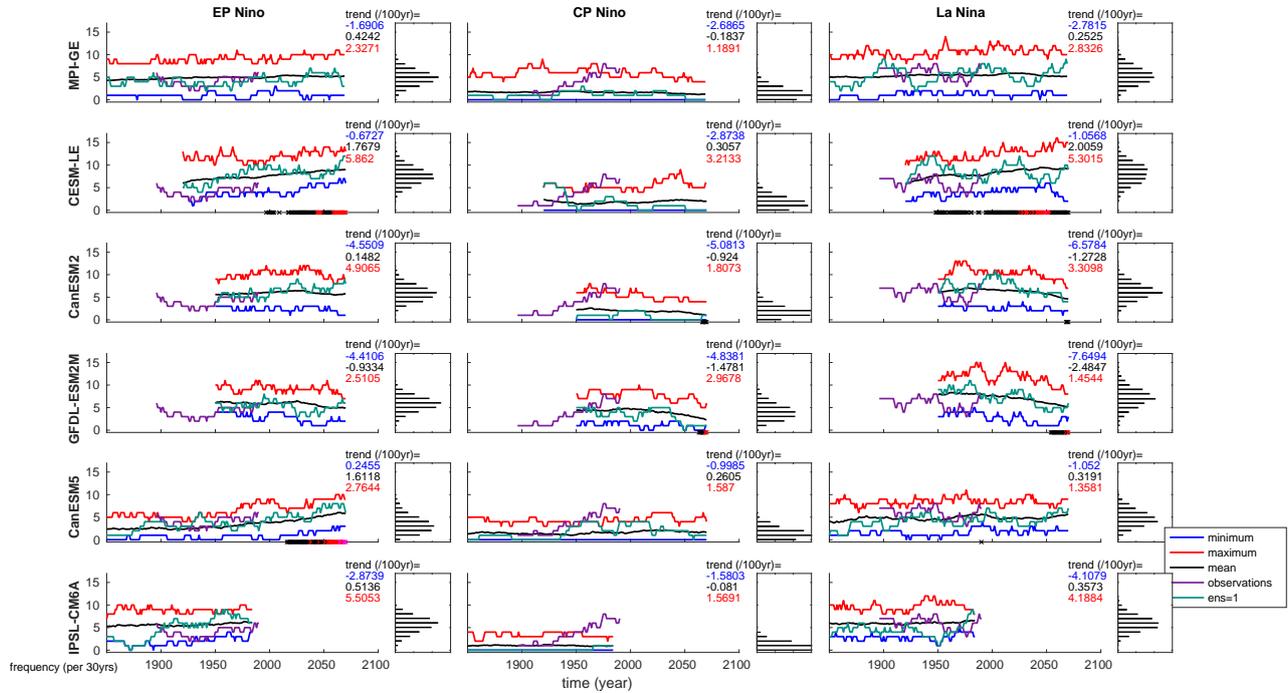


Figure S5. ENSO frequency in each SMILE for EP, CP, and LN events (left, middle and right columns respectively) for the shifted centers of variability. Black line shows the ensemble mean for each year, red line shows the ensemble maximum, the blue line the ensemble minimum, the purple line is HadISST observations, and the green line is the first ensemble member. Frequency is calculated as the number of events in a single ensemble member per 30 years, taken as a running calculation along the time-series. PDFs show the distribution of ensemble members for the entire time-series. Black dots on the x-axis demonstrate when the signal (current ensemble mean minus the ensemble mean at the beginning of the time-series) is greater than the noise (standard deviation taken across the ensemble). Red dots show when the signal is 1.645 times the noise, while magenta dots show the same when the signal is greater than 2 times the noise. These thresholds correspond to the *likely*, *very likely* and *extremely likely* ranges. Maximum (red), mean (black) and minimum (blue) trends across the individual ensemble members are shown in text at the top right of each panel. We note that the trends are calculated over the entirety of the simulation length for each SMILE. This means that due to the different time periods covered, trends are not directly comparable between different SMILES.

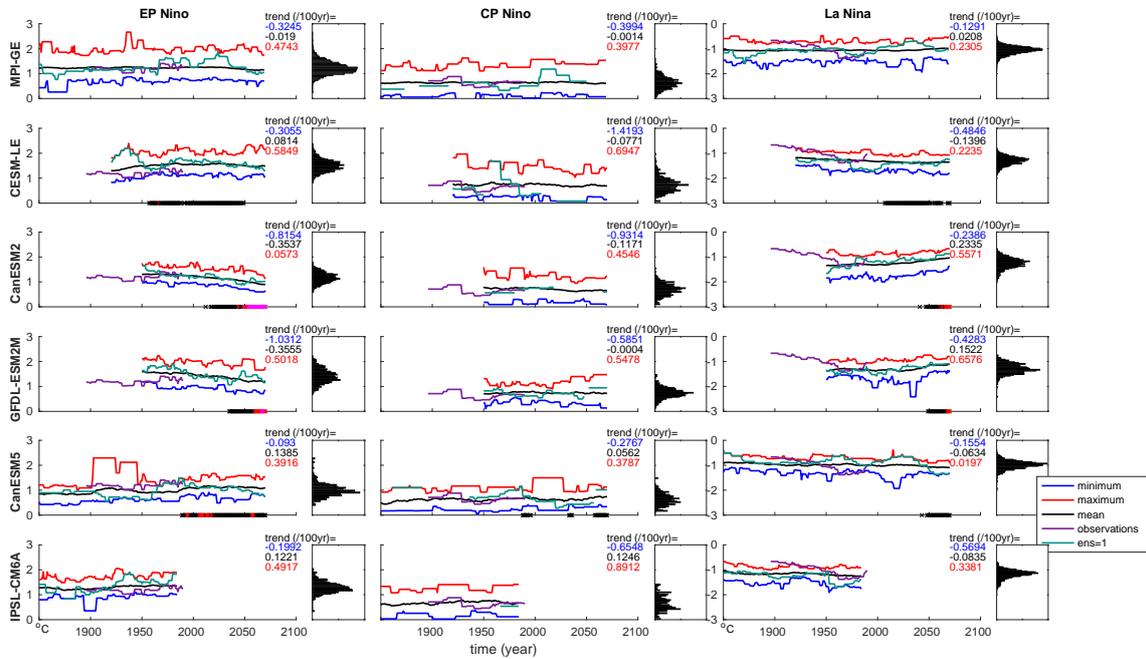


Figure S6. ENSO amplitude in each SMILE for EP, CP and LN events (left, middle and right columns respectively) for the shifted centers of variability. Black line shows the ensemble mean for each year, red line shows the ensemble maximum, the blue line the ensemble minimum, the purple line is HadISST observations, and the green line is the first ensemble member. Amplitude is calculated as the November, December, January mean for the region 160E to 80W between 5N and 5S after the ensemble mean has been removed for each event in a single ensemble member taken as a running calculation along the time-series for 30-year periods. PDFs show the distribution of ensemble members for the entire time-series. Black dots on the x-axis demonstrate when the signal (current ensemble mean minus the ensemble mean at the beginning of the time-series) is greater than the noise (standard deviation taken across the ensemble). Red dots show when the signal is 1.645 times the noise, while magenta dots show the same when the signal is greater than 2 times the noise. These thresholds correspond to the *likely*, *very likely* and *extremely likely* ranges. Maximum (red), mean (black) and minimum (blue) trends across the individual ensemble members are shown in text at the top right of each panel. We note that the trends are calculated over the entirety of the simulation length for each SMILE. This means that due to the different time periods covered, trends are not directly comparable between different SMILEs.

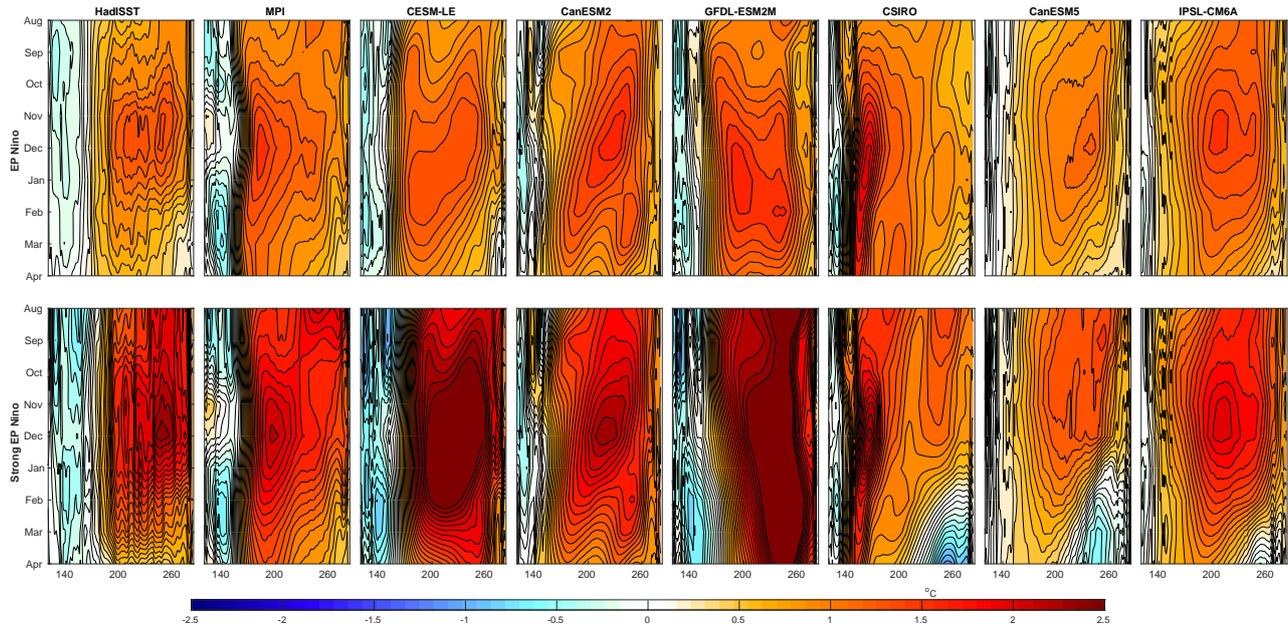


Figure S7. Hovmöller of average SST along the equator in the Pacific Ocean for composites of EP, and ST events (top and bottom row respectively). Shown for HadISST observations (left column) and each individual SMILE (in order of appearance; MPI-GE, CESM-LE, CanESM2, GFDL-ESM2M, CSIRO, CanESM5 and IPSL-CM6A). SST is averaged between 5N and 5S and shown for August to April. SMILE data has the forced response (ensemble mean) removed prior to calculation, HadISST is detrended using a second order polynomial then each months average is removed. The time period used is all of the historical, which is shown for the observations in Table S1 and SMILEs in Table S2.

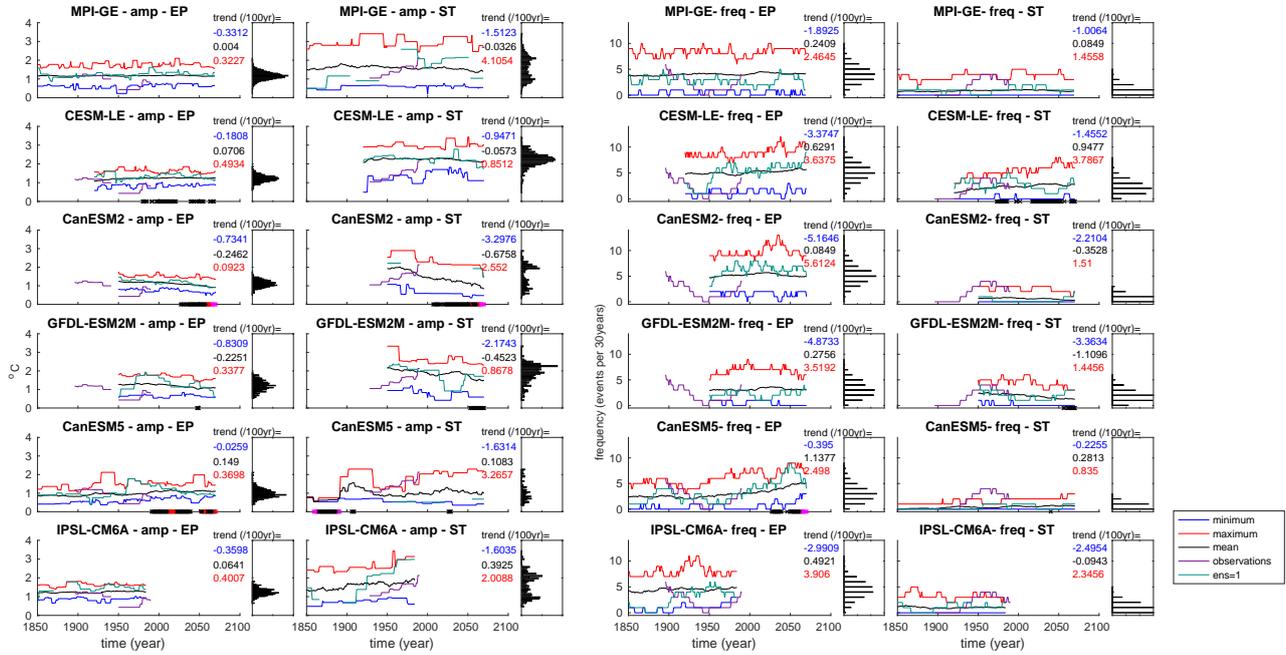


Figure S8. ENSO amplitude and frequency in each SMILE for EP and ST events. Black line shows the ensemble mean for each year, red line shows the ensemble maximum, the blue line the ensemble minimum, the purple line is HadISST observations, and the green line is the first ensemble member. Amplitude is calculated as the November, December, January mean for the region 160E to 80W between 5N and 5S after the ensemble mean has been removed for each event in a single ensemble member taken as a running calculation along the time-series for 30-year periods. Frequency is calculated as the number of events in a single ensemble member per 30 years, taken as a running calculation along the time-series. PDFs show the distribution of ensemble members for the entire time-series. Black dots on the x-axis demonstrate when the signal (current ensemble mean minus the ensemble mean at the beginning of the time-series) is greater than the noise (standard deviation taken across the ensemble). Red dots show when the signal is 1.645 times the noise, while magenta dots show the same when the signal is greater than 2 times the noise. These thresholds correspond to the *likely*, *very likely* and *extremely likely* ranges. Maximum (red), mean (black) and minimum (blue) trends across the individual ensemble members are shown in text at the top right of each panel. We note that the trends are calculated over the entirety of the simulation length for each SMILE. This means that due to the different time periods covered, trends are not directly comparable between different SMILES.

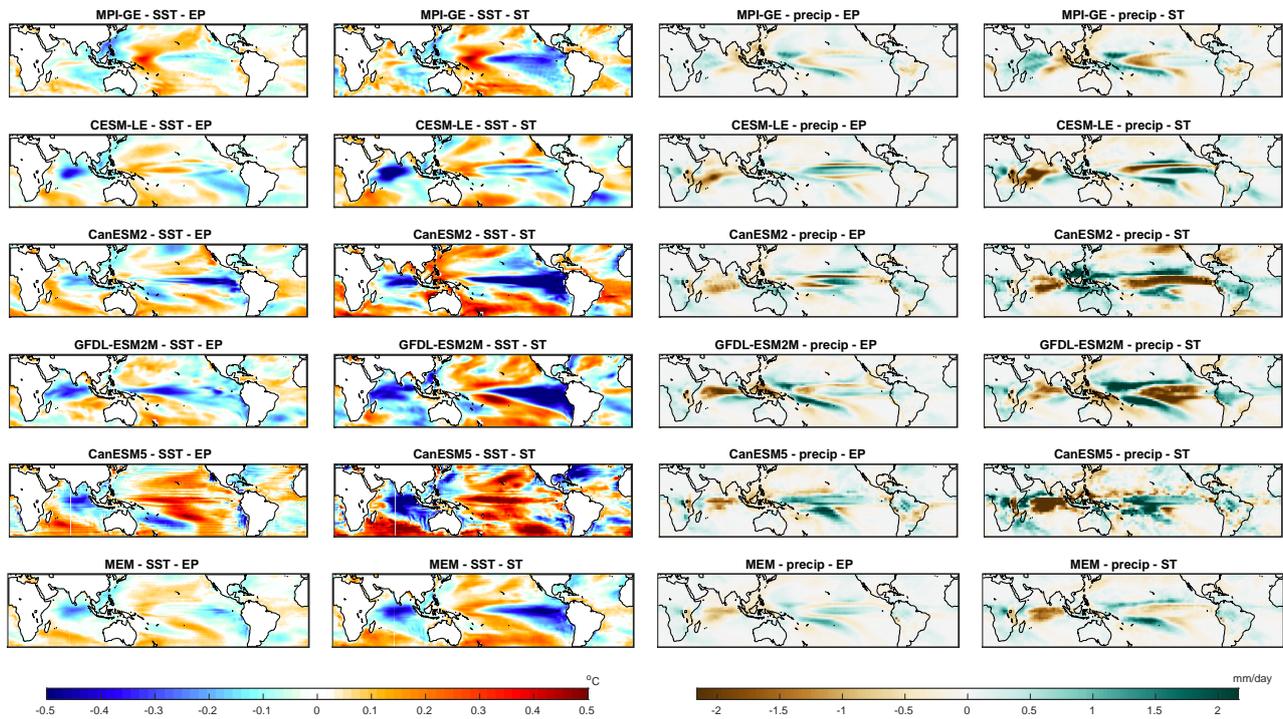


Figure S9. Change in SST and precipitation patterns (left and right columns respectively) in each SMILE in the period 2050-2099 as compared to 1950-1999 for EP and ST events. Shown for each individual SMILE (in order of appearance; MPI-GE, CESM-LE, CanESM2, GFDL-ESM2M, CanESM5 and IPSL-CM5). SST and precipitation patterns are calculated as the November, December, January average and composited for each event type over each time-period. SMILE data has the forced response (ensemble mean) removed prior to calculation.