Supplementary Information to: Large potential of performance-based model weighting to improve decadal climate forecast skill

Vincent Verjans[1*], Markus G Donat[1], Carlos Delgado-Torres[1], Timothy DelSole[2]

[1]Earth Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain.
[2]Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, Virginia, USA.

*Corresponding author(s). E-mail(s): vincent.verjans@bsc.es;

This Supplementary Information includes Figures S1 to S10.

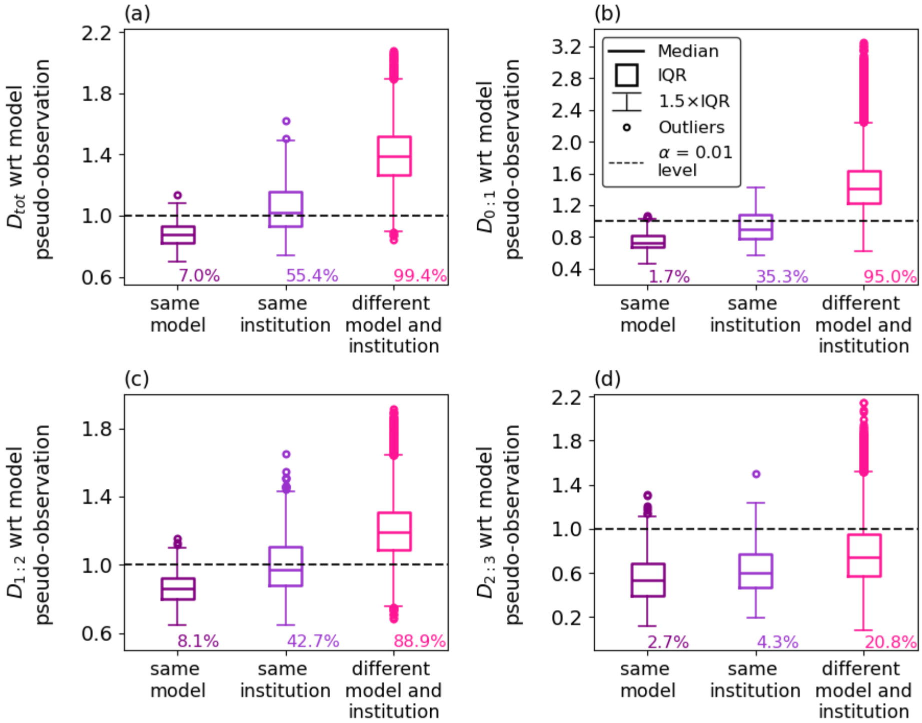

Figure S1. Boxplots of (a) D_{tot}, (b) $D_{0:1}$, (c) $D_{1:2}$, and (d) $D_{2:3}$ values computed from all CMIP6 ensemble members with respect to all the 39 CMIP6 model reference members. The 39 reference members are taken as the first member of each CMIP6 model. Results are grouped separately for members evaluated against a reference from an identical model (same model), from a model developed at the same institution (same institution), and from any other model (different model and institution). Percentages below each boxplot show the proportion of D values significant at the 1% level (horizontal dashed line). Note that the Type 1 error corresponds to an expected proportion of 1% if all members were statistically similar. IQR: inter-quartile range. Panel (a) is identical to Figure 2b in main.

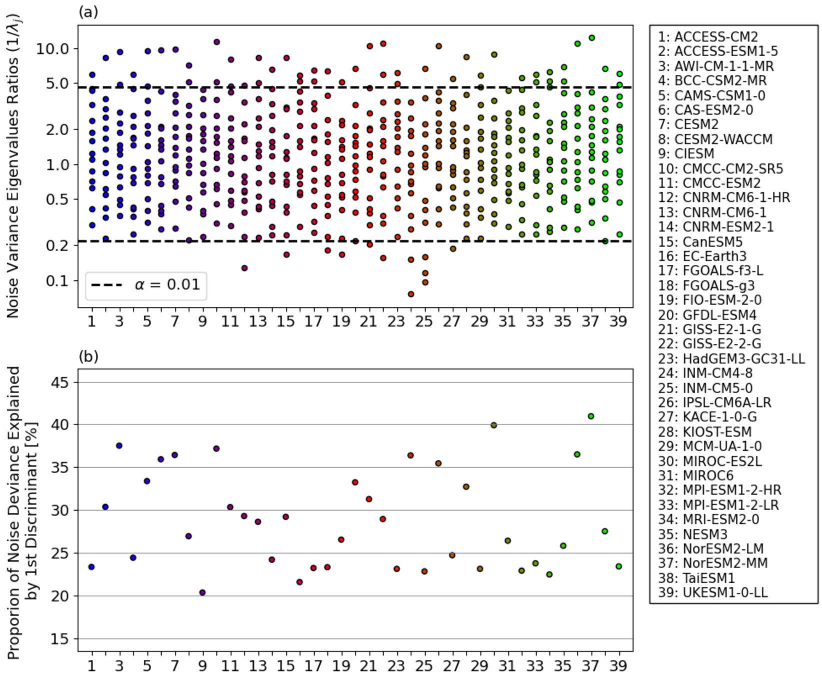

Figure S2. Eigenvalues from covariance discriminant analysis applied to the noise covariance matrices (see Methods in main, Sect. 5.5). (a) The 14 eigenvalues from each model. The horizontal dashed lines show the 0.5% and 99.5% significance thresholds under the hypothesis of equal noise covariance matrix between observations and model members. The eigenvalues are shown for the first member of each model. Note that the inverses of the eigenvalues ($1/\lambda_j$) are shown, such that lower and higher values indicate an under- and over-estimation of the noise magnitude along the corresponding eigenvector direction. (b) The contribution of the first eigenvalue (λ_1) to the noise deviance. Climate model names corresponding to numbers in (a,b) are given in the legend on the right.

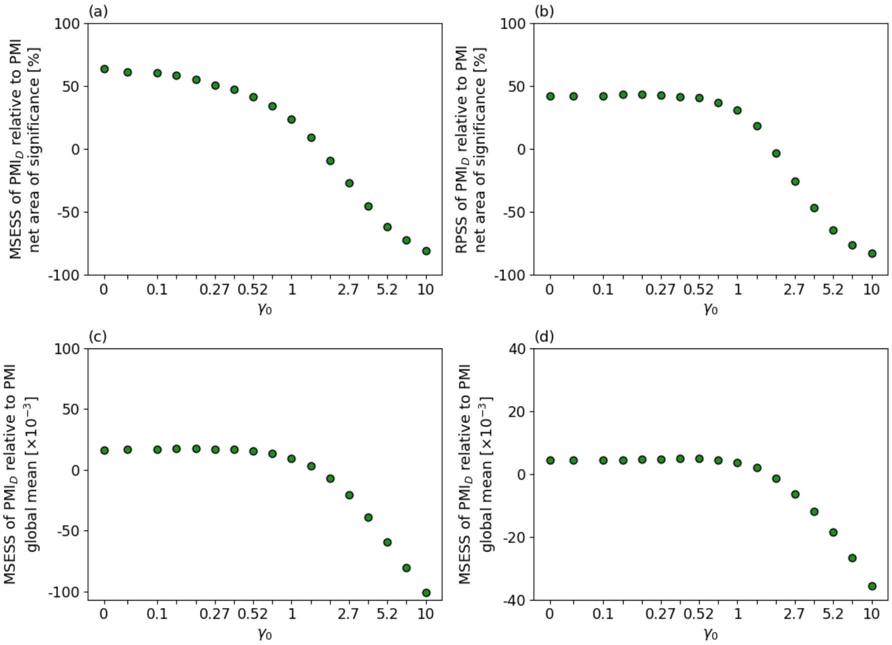

Figure S3. Skill scores of the PMI$_D$ relative to the PMI for pseudo-observation hindcasting. Hindcasts are performed iteratively for all the 316 pseudo-observations from the CMIP6 ensemble (see Table 1 in main). Net area of significant (a) mean squared error skill score (MSESS) and (b) ranked probability skill score (RPSS) of the PMI$_D$ relative to the PMI as reference. Global mean (c) MSESS and (d) RPSS. Both MSESS and RPSS are computed for SST average of hindcast years 1-10, with hindcast start years 1930-2015. Only ensemble members with significant deviance ($D_{tot} > 1$) are available to the model analogue selection procedure to hindcast a given pseudo-observation. Performance differences are shown as a function of γ_0, with fixed $\gamma_1 = \gamma_2 = 1$.

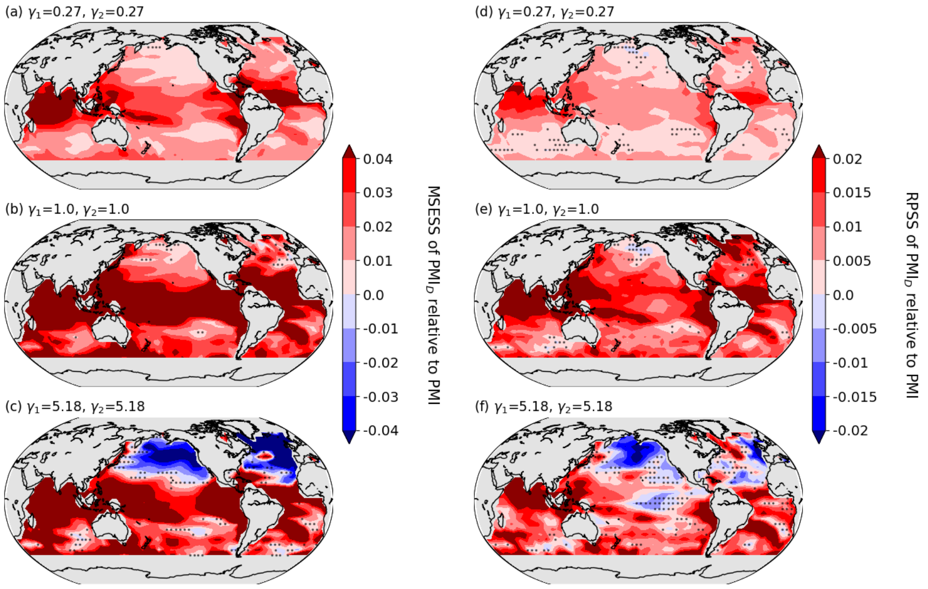
(a) γ_1=0.27, γ_2=0.27
(b) γ_1=1.0, γ_2=1.0
(c) γ_1=5.18, γ_2=5.18
(d) γ_1=0.27, γ_2=0.27
(e) γ_1=1.0, γ_2=1.0
(f) γ_1=5.18, γ_2=5.18
0.04
0.03
0.02
0.01
0.0
-0.01
-0.02
-0.03
-0.04
MSESS of PMI$_D$ relative to PMI
0.02
0.015
0.01
0.005
0.0
-0.005
-0.01
-0.015
-0.02
RPSS of PMI$_D$ relative to PMI

Figure S4. Across-pseudo-observation (a,b,c) mean squared error skill score (MSESS) and (d,e,f) ranked probability skill score (RPSS) of the PMI$_D$ relative to the PMI as reference. The PMI$_D$ is configured at various levels of deviance penalty: low (a,d, $\gamma_1 = \gamma_2 = 0.27$), intermediate (b,e, $\gamma_1 = \gamma_2 = 1$), and high (c,f, $\gamma_1 = \gamma_2 = 5.18$). MSESS and RPSS are computed for SST average of hindcast years 1-10. Skill is aggregated for the 316 pseudo-observations, with hindcast start years 1930-2015 (see Methods in main, Sect. 5.4.1). Stippling indicates absence of significance using a paired data Wilcoxon signed-rank test, while controlling for a false discovery rate $\alpha_{FDR} = 0.1$ (see Methods in main, Sect. 5.4.1). This figure is the same as Figure 6 in main, but without applying the constraint of significant deviance ($D_{tot} > 1$) in the model analogue selection procedure.

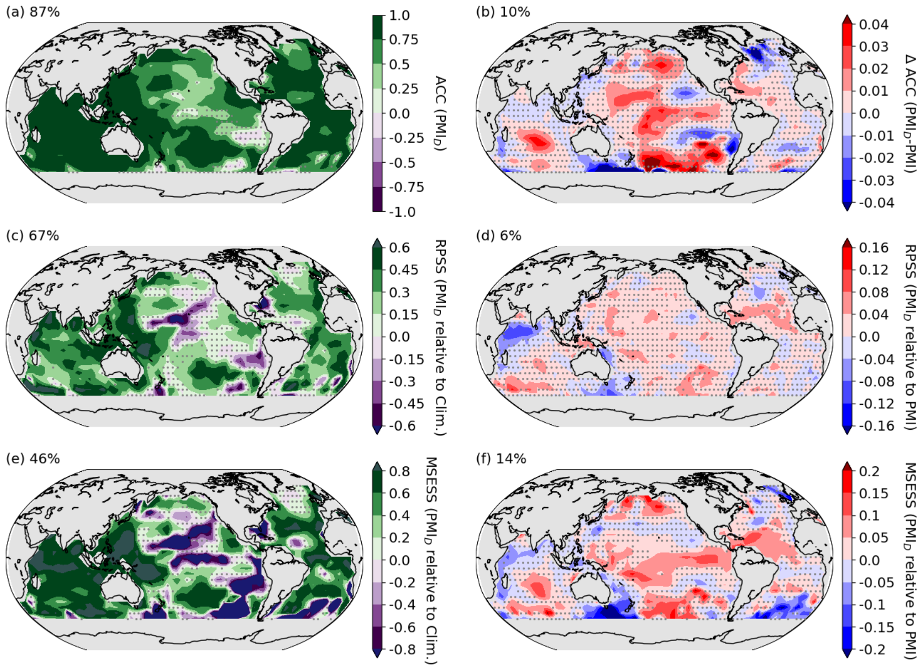

Figure S5. 1930-2024 hindcast performance. Evaluation of predicted SST average of hindcast years 1-10, verified against HadISST. (a) Anomaly corelation coefficient (ACC) of the PMI$_D$. (b) Difference in ACC between the PMI$_D$ and the PMI. Ranked probability skill score (RPSS) of the PMI$_D$ (c) relative to climatology and (d) relative to the PMI as reference. Mean squared error skill score (MSESS) of the PMI$_D$ (e) relative to climatology and (f) relative to the PMI as reference. The total number of 10-year averaged hindcasts is 86, with 1930 and 2015 as first and last start dates. The net percentage of area with significant positive skill is given on top of each map. Stippling indicates absence of significance when controlling for a false discovery rate $\alpha_{FDR} = 0.1$. Note that the deviance values used as penalties in the PMI$_D$ have been computed against the ERSSTv6 reference data set, not HadISST. This figure is the same as Figure 7 in main, but evaluating against HadISST instead of ERSSTv6.

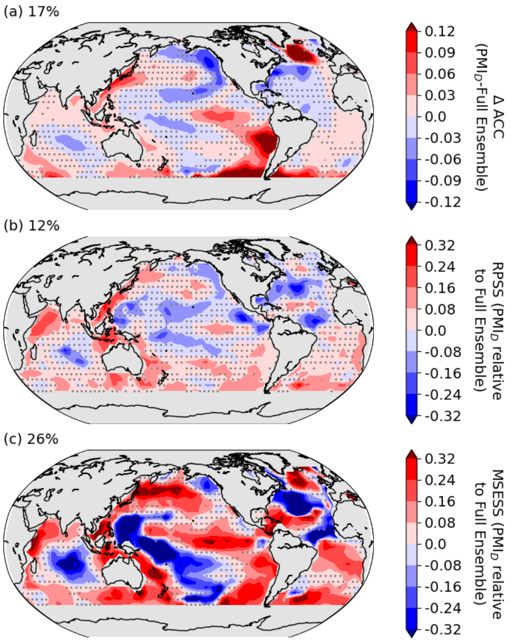

Figure S6. 1930-2024 hindcast performance. Evaluation of predicted SST average of hindcast years 1-10, verified against ERSSTv6. (a) Difference in ACC between the PMI$_D$ and the full CMIP6 ensemble mean. (b) Ranked probability skill score (RPSS) relative to the full CMIP6 ensemble as reference. (c) Mean squared error skill score (MSESS) computed relative to the full CMIP6 ensemble mean as reference. The total number of 10-year averaged hindcasts is 86, with 1930 and 2015 as first and last start dates. The net percentage of area with significant positive skill is given on top of each map. Stippling indicates absence of significance when controlling for a false discovery rate $\alpha_{FDR} = 0.1$.

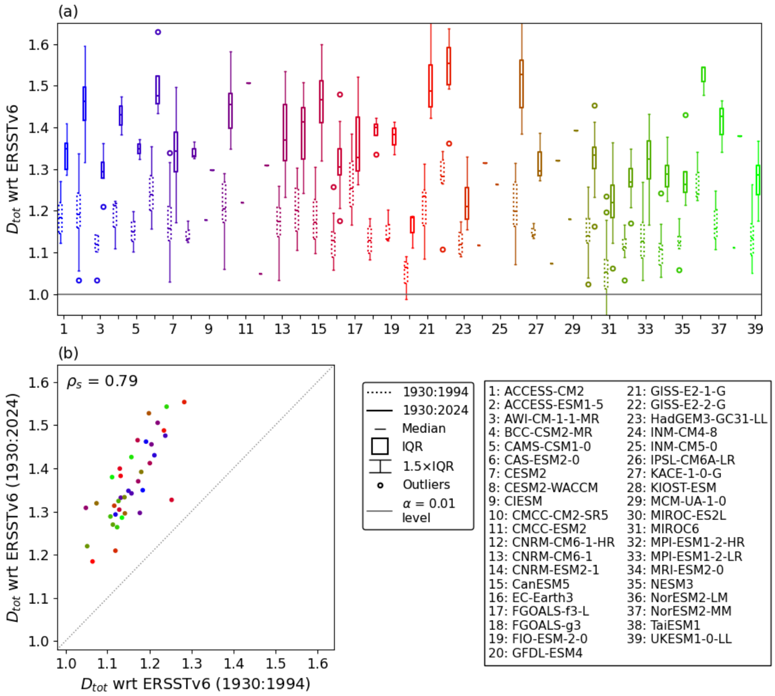

Figure S7. Correspondence between total deviance (D_{tot}) evaluated with respect to 1930-1994 and 1930-2024 ERSSTv6 observational reference. (a) Boxplots of D_{tot} for 1930-1994 (dotted) and 1930-2024 (solid) periods, showing spread from different members. (b) Scatterplot of ensemble median D_{tot} of each model in 1930-1994 versus 1930-2024. The Spearman rank correlation coefficient (ρ_s) is given and highly significant (p-value $< 10^{-8}$). Higher D_{tot} indicates stronger statistical inconsistency with ERSSTv6, with the value 1 corresponding to statistical significance at the 1% level (horizontal line in a). Note that the shift towards higher D_{tot} for the longer reference period is expected, as more data provides stronger evidence for any deviance between modeled and observed SST. Climate model names corresponding to numbers in (a) are given in the legend on the right. IQR: inter-quartile range.

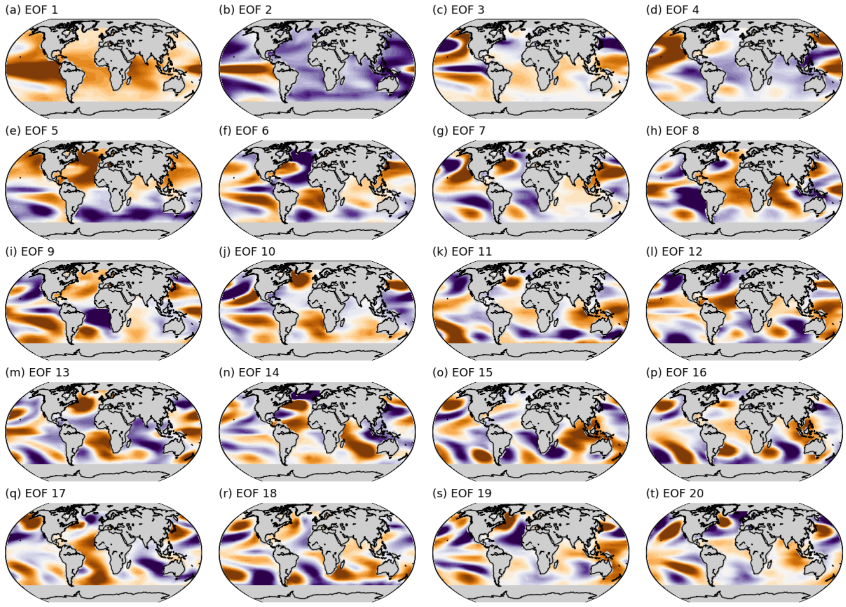

Figure S8. The 20 leading Empirical Orthogonal Functions computed from the CMIP6 historical simulations. To perform principal component analysis, we use one ensemble member of each one of the 39 climate models is taken (see Table 1 in main). The total variance of each ensemble member is scaled to unity, and members are concatenated along the time dimension. Note that SST data are annually-averaged from July to June (see Methods in main, Sect. 5.2.1).

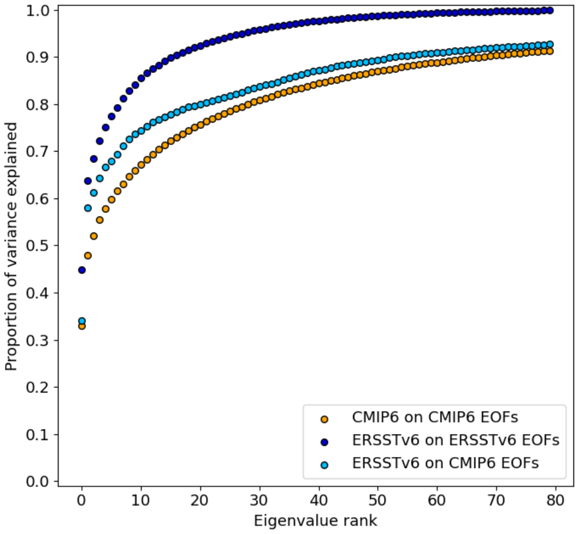

Figure S9. Variance explained through projection of SST data on an increasing number of Empirical Orthogonal Functions (EOFs). The orange curve shows the proportion of variance for the CMIP6 ensemble (one historical simulation per model) when projected on the CMIP6 ensemble EOFs. The light blue curve shows the proportion of variance for the ERSTv6 data when projected on the CMIP6 ensemble EOFs. The dark blue curve shows the proportion of variance for the ERSSTv6 data when projected on the EOFs specific to the ERSSTv6 SST data. Note that SST data are annually-averaged from July to June (see Methods in main, Sect. 5.2).

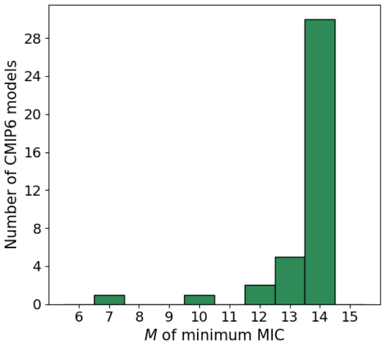

Figure S10. Optimal number M of Empirical Orthogonal Functions included in the VARX modeling framework to minimize the Mutual Information Criterion (MIC). The MIC formulation is given in the Methods in main, Sect. 5.2.1. Note that the MIC optimization is performed on the SST of the historical simulation of the first ensemble member of each of the 39 CMIP6 model (see Table 1 in main).