

Introduction & Motivation

- Decadal (1-10 yr) climate change is the combined result of a forced component and of the internal variability intrinsic to the climate system.
- The poor man's initialization (PMI) selects the model analogues from large ensembles of projections that best match the observed initial state at forecast start time.
- We extend the PMI to include a performance-based model weight in the selection of analogues, favoring models statistically consistent with the observed SST record.
- We validate the model weighting in pseudo-observation SST hindcasts, test its skill in real-world hindcasts, and highlight challenges in performance-based model weighting.

Methods: the deviance statistic

- The deviance is a measure of statistical consistency between observed (o) and model-simulated (m) SST spatio-temporal variability. (observations: ERSSTv6 1930-2024)
- Ideally, model output and observational records are statistically indistinguishable.

$$\begin{cases} \mathbf{y}_o(t) = \mathbf{c}_o + \mathbf{B}_o^T \mathbf{y}_o(t-1) + \mathbf{a}_o z(t) + \boldsymbol{\epsilon}_o(t), & \boldsymbol{\epsilon}_o \sim \boldsymbol{\Sigma}_o \\ \mathbf{y}_m(t) = \mathbf{c}_m + \mathbf{B}_m^T \mathbf{y}_m(t-1) + \mathbf{a}_m z(t) + \boldsymbol{\epsilon}_m(t), & \boldsymbol{\epsilon}_m \sim \boldsymbol{\Sigma}_m \end{cases}$$

\mathbf{y} : SST principal components $\boldsymbol{\Sigma}$: annual noise covariance matrix
 \mathbf{B} : linearized dynamics \mathbf{a} : radiative forcing response
 \mathbf{c} : constant z : radiative forcing (input)

Nested hypotheses:

- Ω_0 : $\boldsymbol{\Sigma}_o, \boldsymbol{\Sigma}_m$ unrestricted, $\mathbf{B}_o, \mathbf{B}_m$ unrestricted, and $\mathbf{a}_o, \mathbf{a}_m$ unrestricted
- Ω_1 : $\boldsymbol{\Sigma}_o = \boldsymbol{\Sigma}_m$, $\mathbf{B}_o, \mathbf{B}_m$ unrestricted, and $\mathbf{a}_o, \mathbf{a}_m$ unrestricted
- Ω_2 : $\boldsymbol{\Sigma}_o = \boldsymbol{\Sigma}_m$, $\mathbf{B}_o = \mathbf{B}_m$ and $\mathbf{a}_o, \mathbf{a}_m$ unrestricted
- Ω_3 : $\boldsymbol{\Sigma}_o = \boldsymbol{\Sigma}_m$, $\mathbf{B}_o = \mathbf{B}_m$, $\mathbf{a}_o = \mathbf{a}_m$

- Ω_0 vs. $\Omega_1 \rightarrow D_{0:1}$: noise deviance
- Ω_1 vs. $\Omega_2 \rightarrow D_{1:2}$: dynamical deviance
- Ω_2 vs. $\Omega_3 \rightarrow D_{2:3}$: radiative forcing response deviance
- $D_{tot} = D_{0:1} + D_{1:2} + D_{2:3}$: total deviance

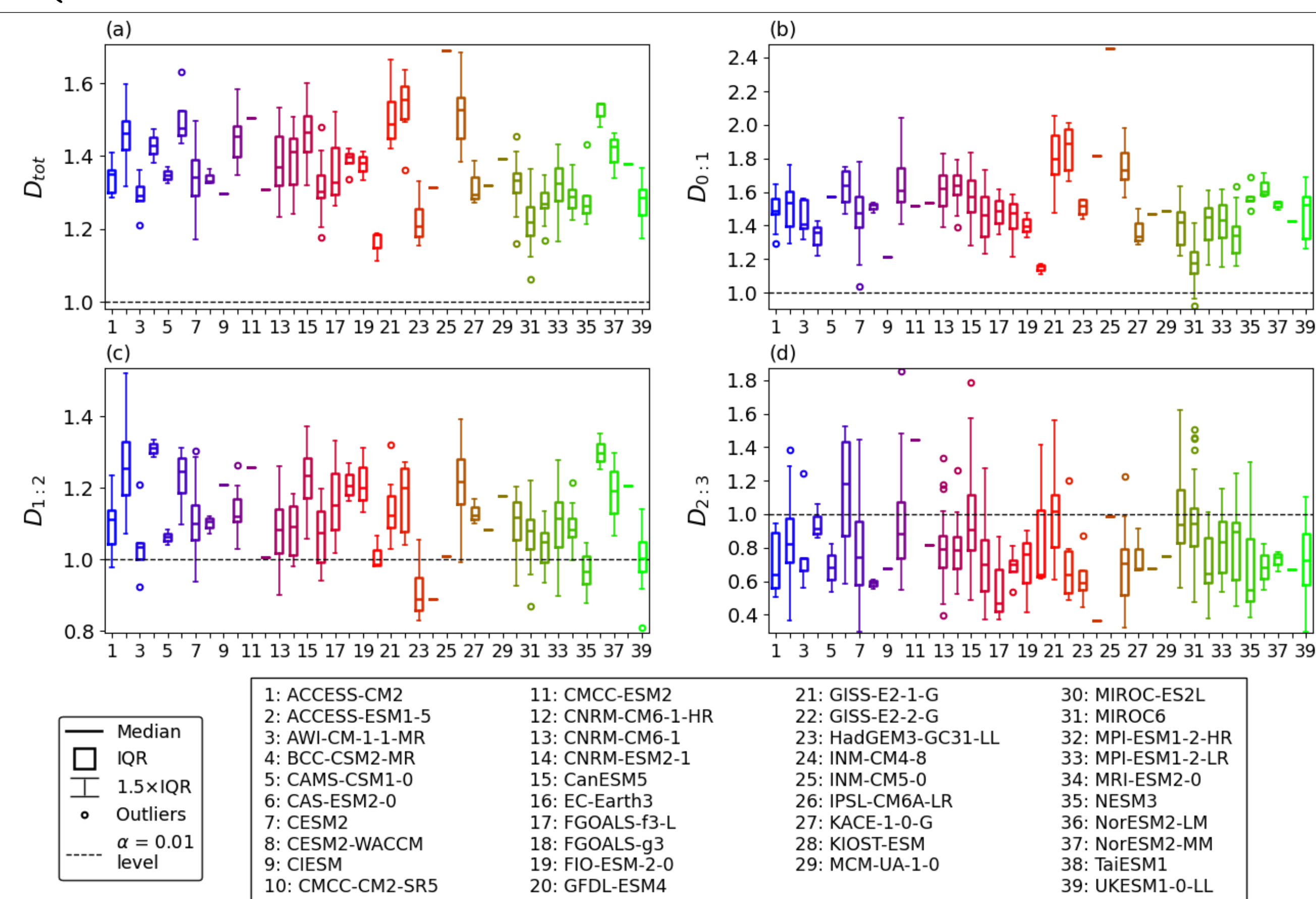
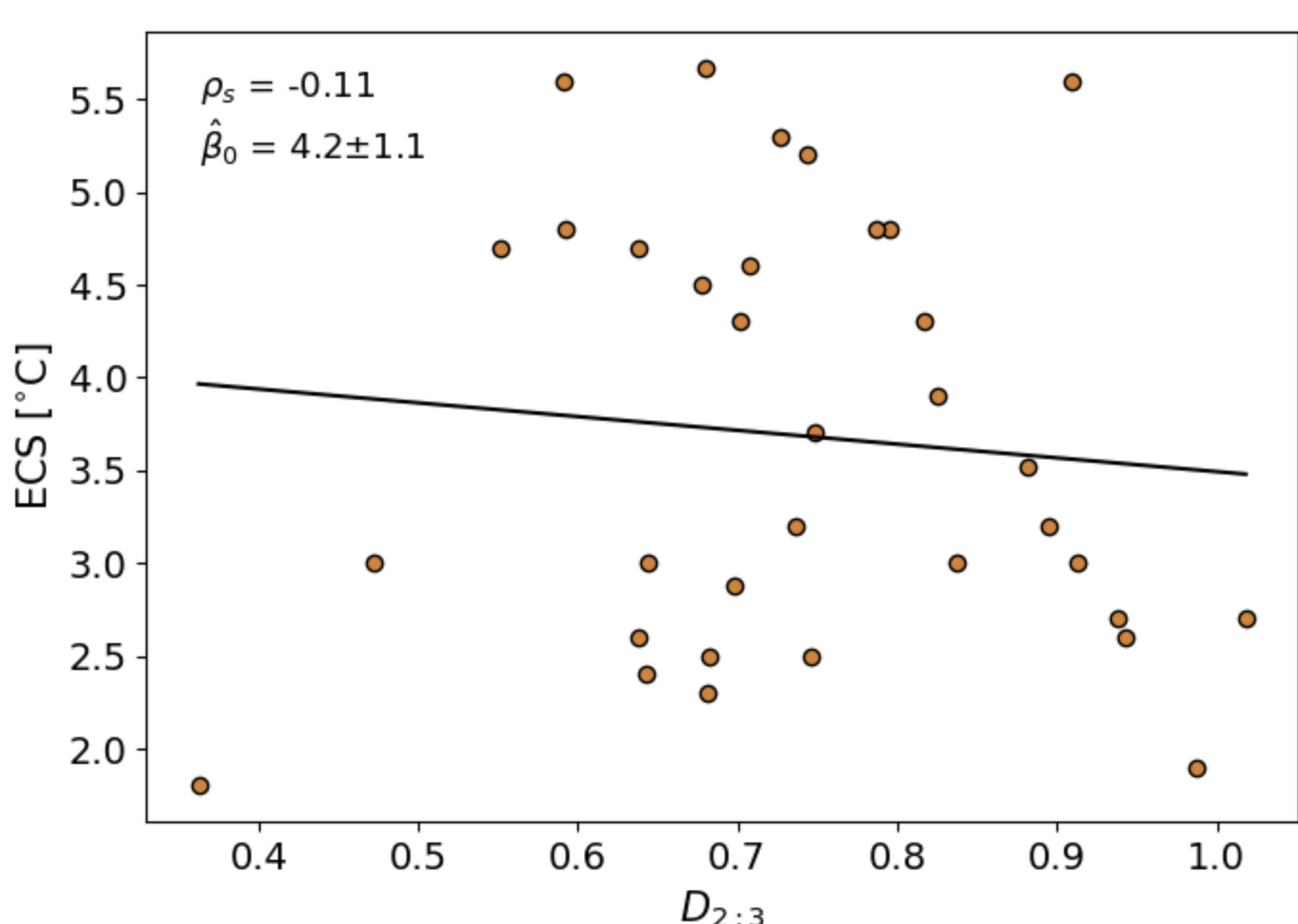


Figure 1: SST deviance diagnostic. (a) Total deviance, (b) noise deviance, (c) dynamical deviance, and (d) radiative forcing response deviance from 39 CMIP6 models with respect to ERSSTv6 (1930-2024) observational reference. Boxplots show spread from different members. Higher D indicates stronger statistical inconsistency with ERSSTv6, with the value of 1 corresponding to statistical significance at the 1% level (dashed horizontal lines).

Connection to Equilibrium Climate Sensitivity?



- Since $D_{2:3}$ compares observed versus model response to radiative forcing, we investigate its relation to ECS.
 - Slight evidence for higher ECS range.

Figure 3: Relation between Equilibrium Climate Sensitivity and deviance in radiative forcing response. CMIP6 model ensemble median $D_{2:3}$ is shown versus its ECS (Meehl et al., 2020). Spearman rank correlation coefficient (ρ_s) is given, but not significant ($p > 0.1$). Intercept of the linear fit (β_0) is given with its standard error.

Critical challenges in validating performance-based model weighting

- Similarity between unweighted and weighted ensembles.
 - Observational record provides few independent verification samples of decadal climate variability.
 - There is large sampling variability in predictive skill, amplified by autocorrelation inherent to SST.
 - Limited inherent predictability reduces potential skill improvements from model weighting.
 - Tradeoffs: model weighting may come at a cost. Here, we sacrifice some initial state consistency.
- Over many hindcasts, we show a robust skill improvement from model weighting:
 - Highly significant skill improvement in the pseudo-observation hindcasts (Figs. 2, 3)
 - Increase in skill with number of real-world hindcasts (Fig. 5)
 - The potential improvement in forecast skill is large. However, it is not guaranteed: for any individual forecast, the outcome is subject to high variability.

Pseudo-observation hindcasts

- Multi-model ensemble of 316 members. We predict each member, using the remaining 315 as potential analogues.
- 1-10 yr hindcasts with 86 start dates (1930-2015).
- Analogue selection at any hindcast start date t :
 - The PMI finds the 30 members minimizing previous 9-year mean SST RMSE
 - We introduce the $PMI_D \Rightarrow$ includes a deviance penalty

$$s_k(t) = \frac{1}{\text{RMSE}(t)_k} - \gamma_0 D_{0:1,k} - \gamma_1 D_{1:2,k} - \gamma_2 D_{2:3,k} \Rightarrow \text{score of member } k \text{ at start date } t$$

- We calibrate the $(\gamma_0, \gamma_1, \gamma_2)$ combination by optimizing MSESS and RPSS across the 316 pseudo-observation hindcasts ($316 \times 86 = 27,179$ decadal hindcasts).
- Can the PMI_D improve the decadal skill of the PMI?

Figure 2: Global skill gain in pseudo-observation hindcasts. Net area of significant (a) mean squared error skill score (MSESS) and (b) ranked probability skill score (RPSS) of the PMI_D against the reference PMI. Global mean (c) MSESS and (d) RPSS. Skill is computed for predicted SST average of hindcast years 1-10. Hatching shows non-significance at 10% level from bootstrap procedure.

Note: results are insensitive to γ_0 , so we fix it at 0.

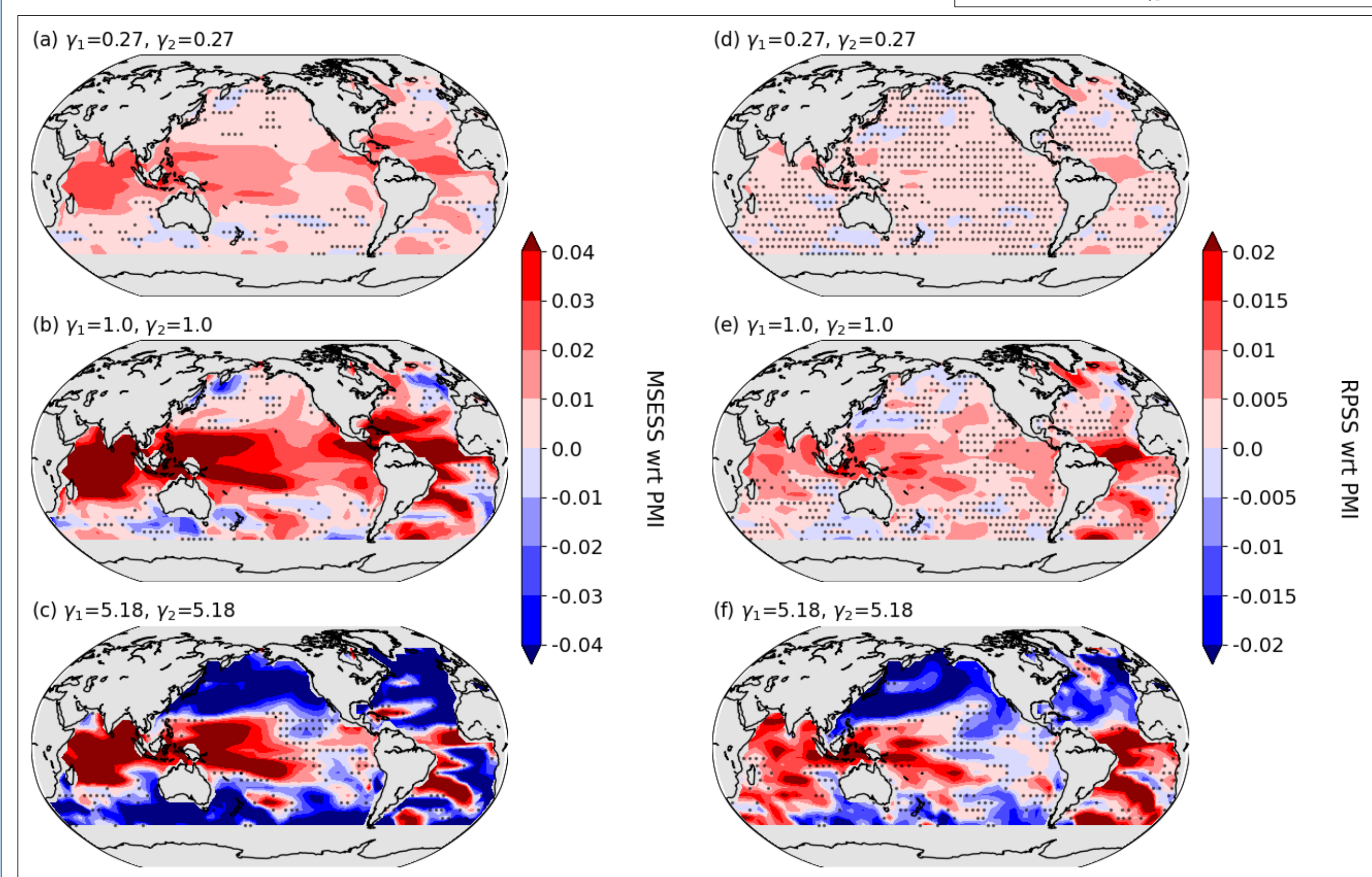
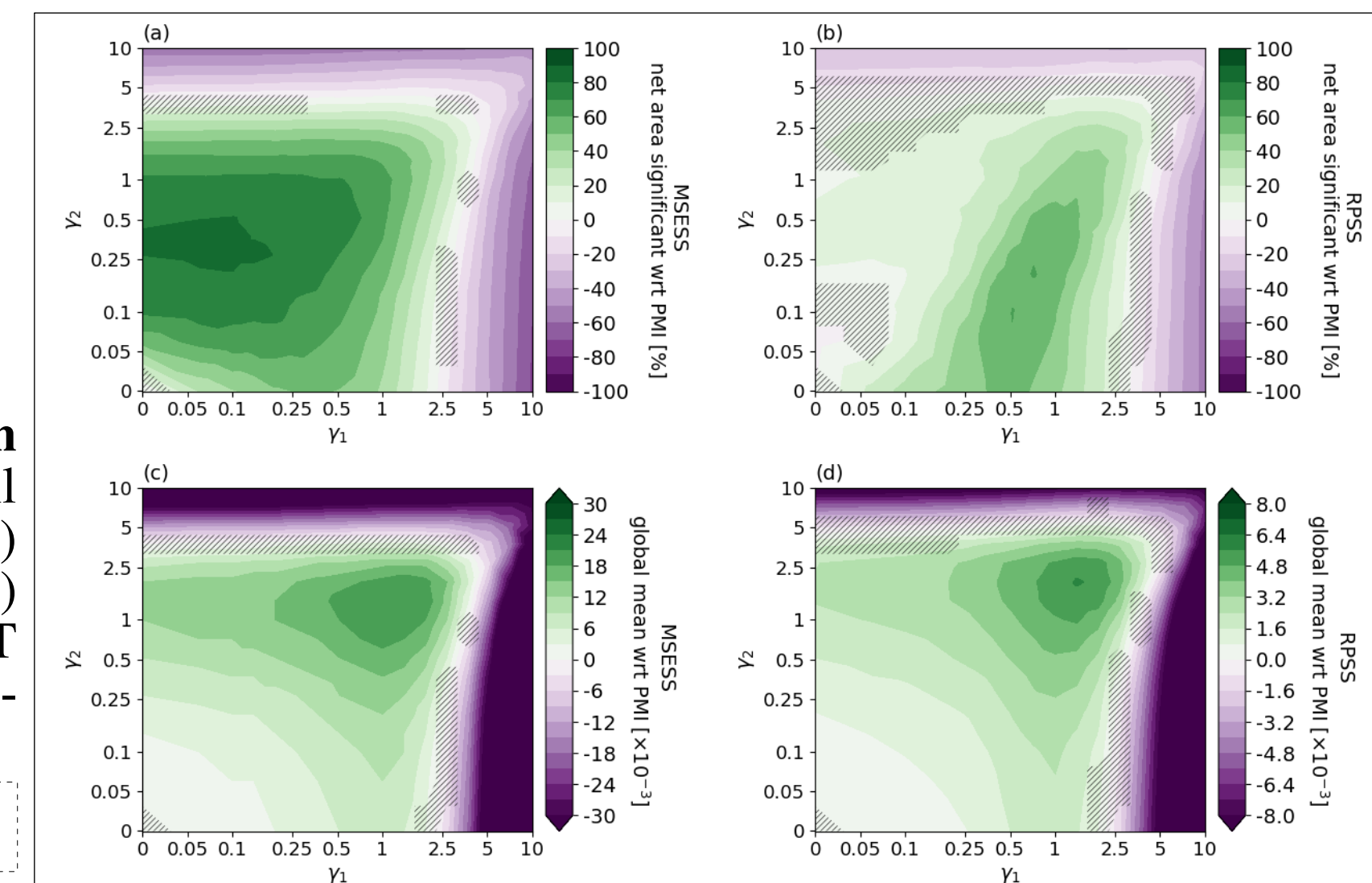


Figure 3: Spatial skill gain in pseudo-observation hindcasts. (a,b,c) MSESS and (d,e,f) RPSS of the PMI_D against the reference PMI. Skill change at (a,d) low, (b,e) intermediate, and (c,f) high levels of deviance penalty. Skill is for SST average of hindcast years 1-10. Stippling shows non-significance, controlling for a false discovery rate of 10%.

Real-world hindcasts

- 1-10 yr hindcast verification with ERSSTv6 (start dates 1930-2015). Comparison of PMI_D and PMI skill.
- PMI_D is significantly more skillful than PMI over 8% (ACC) and 5% (MSESS) of the oceans, but -1% (RPSS).
- First observation-based verification of skill improvement from performance-based model weighting.

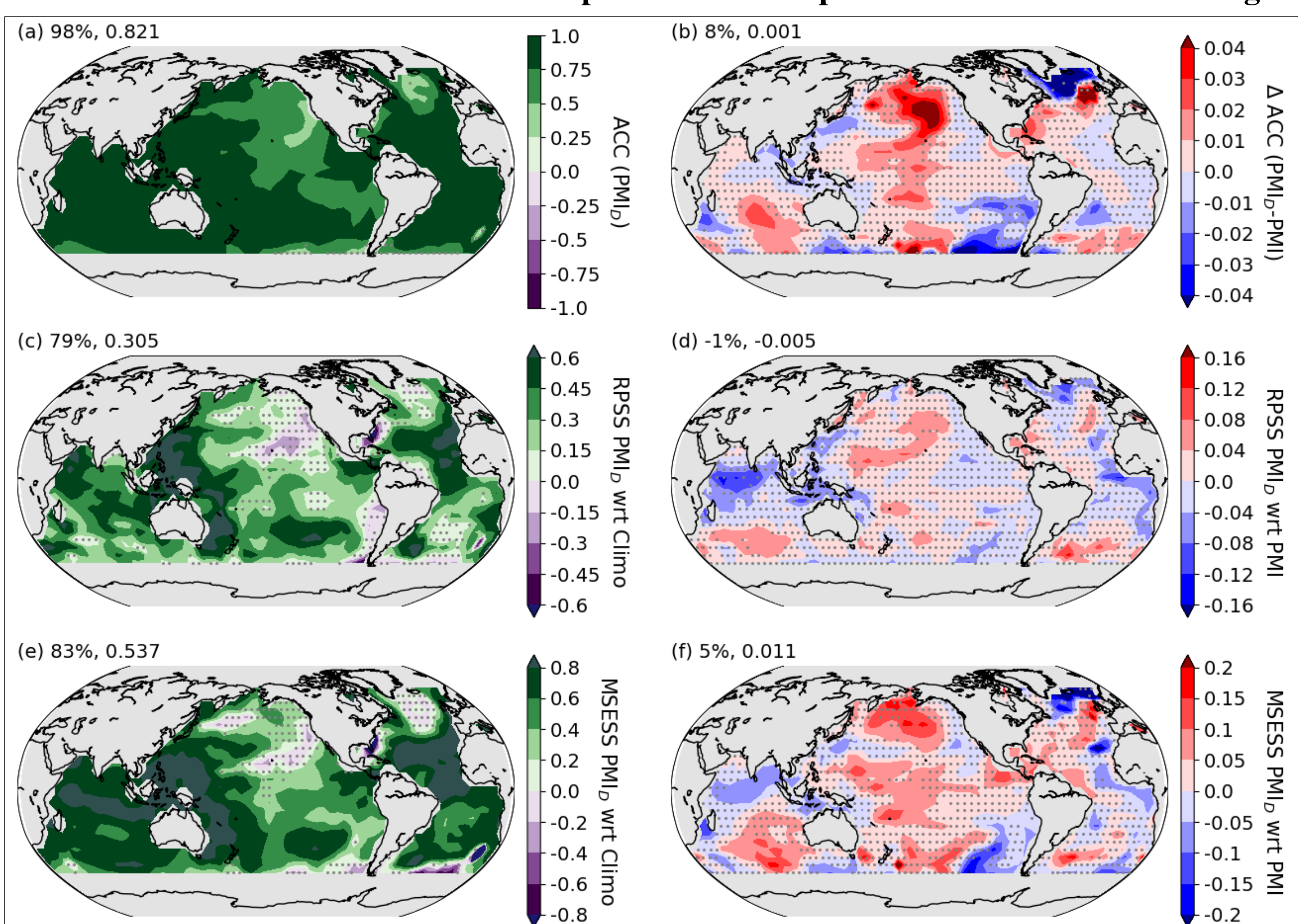


Figure 5: Spatial skill gain in real-world hindcasts. (a) Anomaly correlation coefficient (ACC), (b) difference in ACC wrt PMI, RPSS against reference of (c) climatology and (d) PMI, and MSESS against reference of (e) climatology and (f) PMI. Skill is for SST average of hindcast years 1-10, for 86 start dates (1930-2015). Verification data set: ERSSTv6. Stippling shows non-significance, controlling for a false discovery rate of 10%. Numbers on top of maps give net % area of significant skill gain, and global mean.

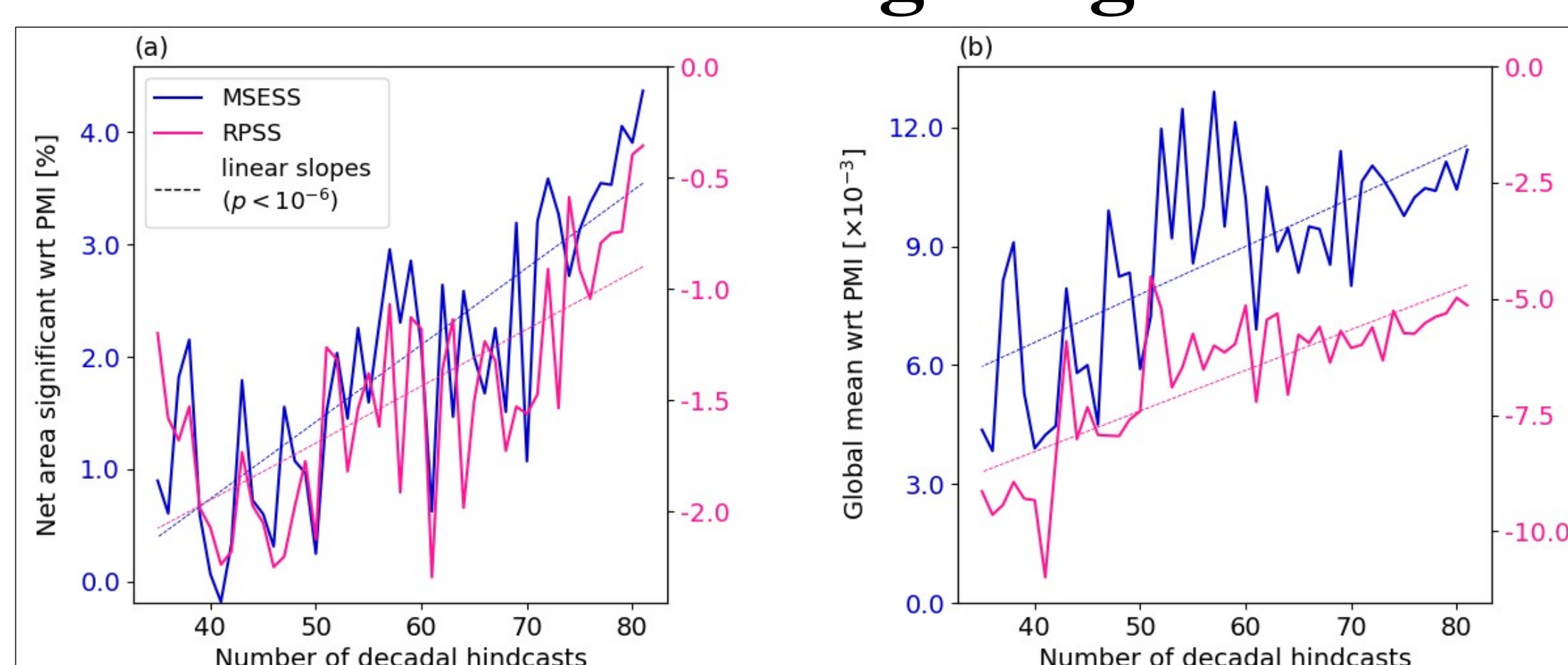


Figure 6: Dependence of PMI_D skill gain on number of hindcasts. Skill change relative to PMI as the original number of verification decadal hindcasts (86) is reduced. For each subset of hindcasts, 30 continuous periods are left-out from the 1930-2015 period. The mean skill gain across these 30 samples is shown.

UPCLIV Workshop 2025
Bologna, Italy

Funded by the European Union
Horizon Europe
EXPECT project
Grant 101137636

