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

Verjans, V.[1], Donat, M.[1], Delgado-Torres, C.[1], and DelSole, T.[2]

1: Earth Sciences Department, Barcelona Supercomputing Center, Spain

2: Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, USA

vincent.verjans@bsc.es

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 analogue-based initialization (ABI) selects the model analogues from large ensembles of projections that best match the observed initial state at forecast start time.
- We extend the ABI 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.

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

$\boldsymbol{y}$: SST principal components — $\boldsymbol{\Sigma}$: annual noise covariance matrix

$\mathbf{B}$: linearized dynamics — $\boldsymbol{a}$: radiative forcing response

$\boldsymbol{c}$: constant — z : radiative forcing (input)

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

$$\begin{cases} \Omega_0 \text{ vs. } \Omega_1 \rightarrow D_{0:1} : \text{noise deviance} \\ \Omega_1 \text{ vs. } \Omega_2 \rightarrow D_{1:2} : \text{dynamical deviance} \\ \Omega_2 \text{ vs. } \Omega_3 \rightarrow D_{2:3} : \text{radiative forcing response deviance} \\ D_{tot} = D_{0:1} + D_{1:2} + D_{2:3} : \text{total deviance} \end{cases}$$

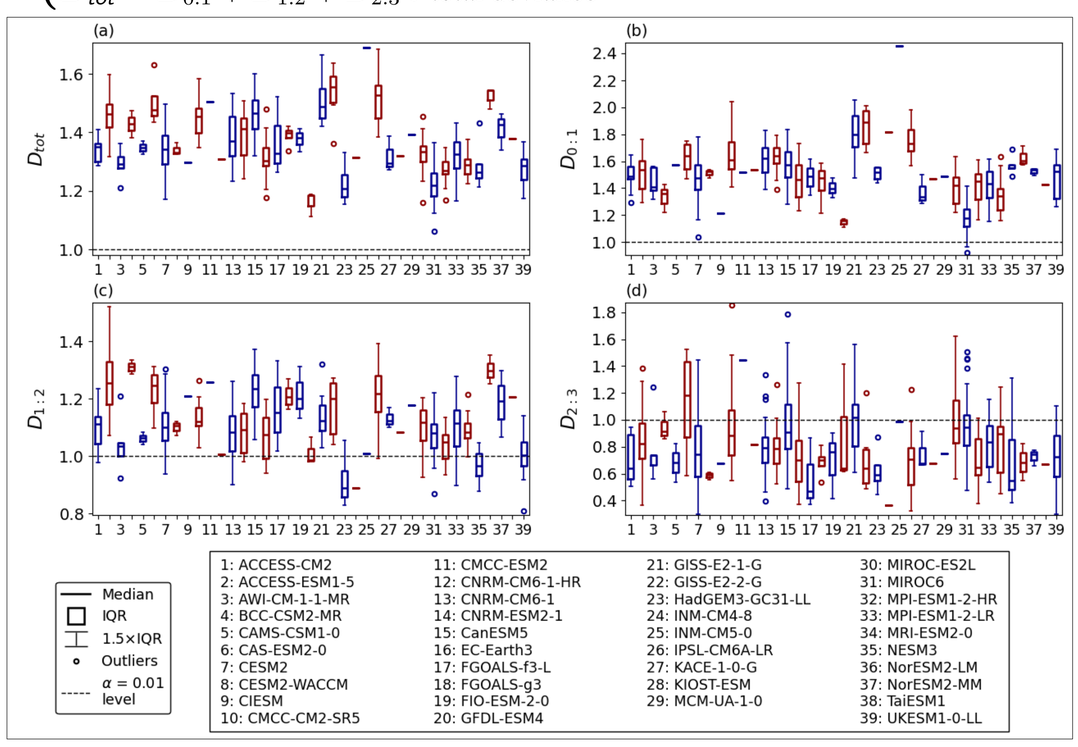

Figure 1: SST deviance diagnostic. (a) Total deviance, (b) noise deviance, (c) dynamical deviance, and (d) radiative forcing response deviance from 39 CMIP6 models with respect to ERSSTv6 (1930-2024) observational reference. Boxplots show spread from different members. Higher D indicates stronger statistical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

Critical challenges in validating performance-based model weighting

(1) **Similarity between unweighted and weighted ensembles**.
(2) **Observational record provides few independent verification samples of decadal climate variability**.
(3) **There is large sampling variability in predictive skill, amplified by autocorrelation inherent to SST**.
(4) **Limited inherent predictability reduces potential skill improvements from model weighting**.
(5) **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:
  (a) Highly significant skill gain in the pseudo-observation hindcasts (Figs. 2, 3)
  (b) 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.

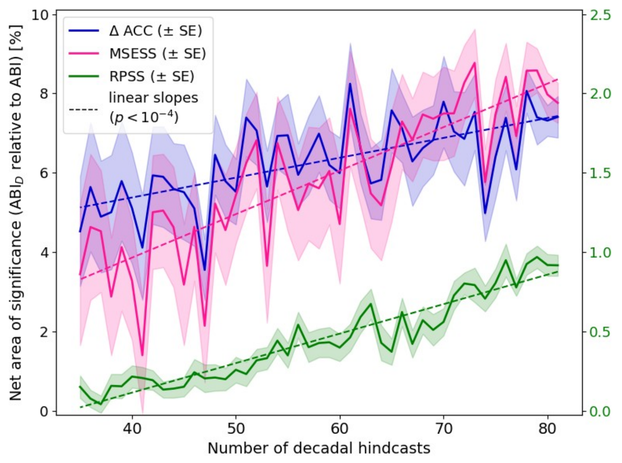

Figure 6: Dependence of ABI$_D$ skill gain on number of hindcasts. Skill change relative to ABI as the original number of verification decadal hindcasts (86) is reduced. For each subset of hindcasts, 30 continuous periods are left-out from the 1930-2015 period. The mean skill gain and its standard error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

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:
  (1) The ABI finds the 30 members minimizing previous 9-year mean SST RMSE
  (2) We introduce the ABI$_D$ ⇒ includes a deviance penalty

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

- We calibrate the (γ_0, γ_1, γ_2) combination by optimizing MSESS and RPSS across the 316 pseudo-observation hindcasts (316 × 86 = 27,179 decadal hindcasts).
- **Can the ABI$_D$ improve the decadal skill of the ABI?**

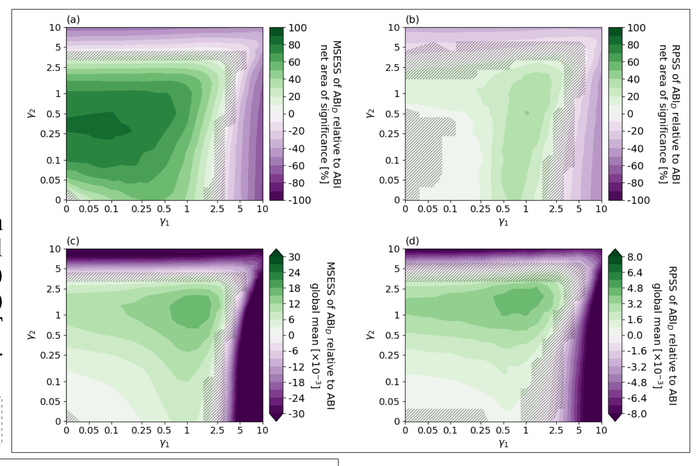

Figure 2: Global skill gain in pseudo-observation hindcasts. Net area of significant (a) mean squared error skill score (MSESS) and (b) ranked probability skill score (RPSS) of the ABI$_D$ against the reference ABI. Global mean (c) MSESS and (d) RPSS. Skill is computed for predicted SST average of hindcast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

Note: results are insensitive to γ_0 ; therefore, we fix γ_0=0.

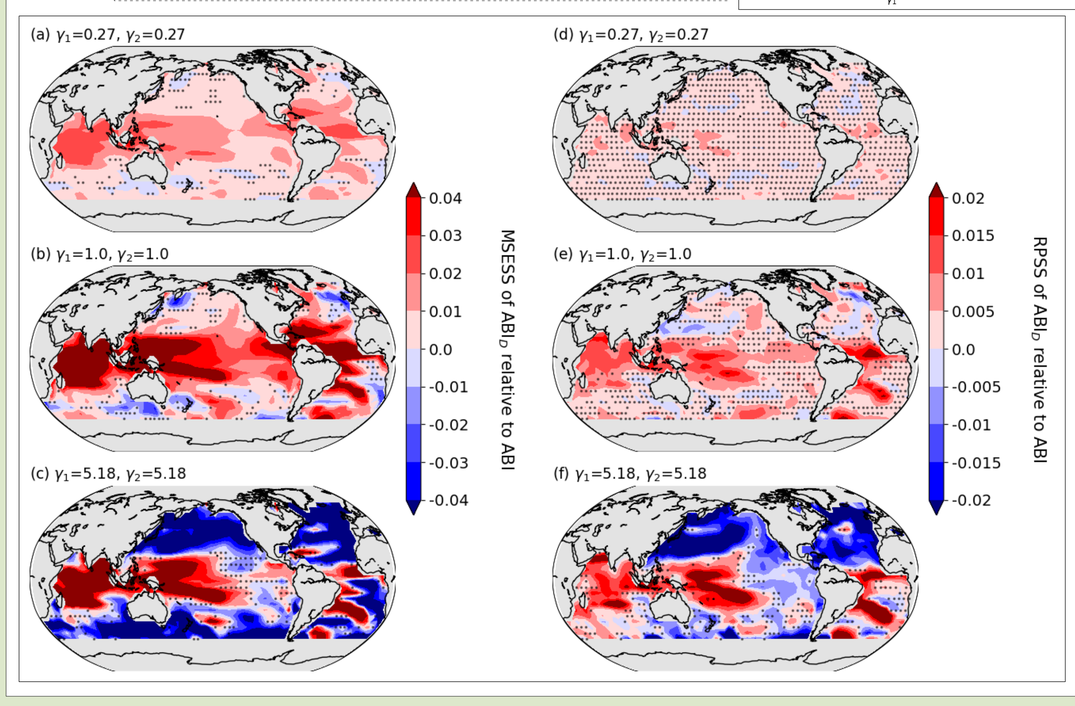

Figure 3: Spatial skill gain in pseudo-observation hindcasts. (a,b,c) MSESS and (d,e,f) RPSS of the ABI$_D$ against the reference ABI. Skill change at (a,d) low, (b,e) intermediate, and (c,f) high levels of deviance penalty. Skill is for SST average of hindcast years 1-10. Stippling shows non-significance,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

Real-world hindcasts

- 1-10 yr hindcast verification with ERSSTv6 (start dates 1930-2015). Comparison of ABI$_D$ and ABI skill.
- ABI$_D$ is significantly more skillful than ABI over 8% (ACC), 1% (RPSS) and 5% (MSESS) of the oceans.
- ABI$_D$ on par with DCPP: skill is significantly larger over 1% (ACC) and 4% (MSESS), but lower over 5% (RPSS).
- **First observation-based verification of skill improvement from performance-based model weighting**.

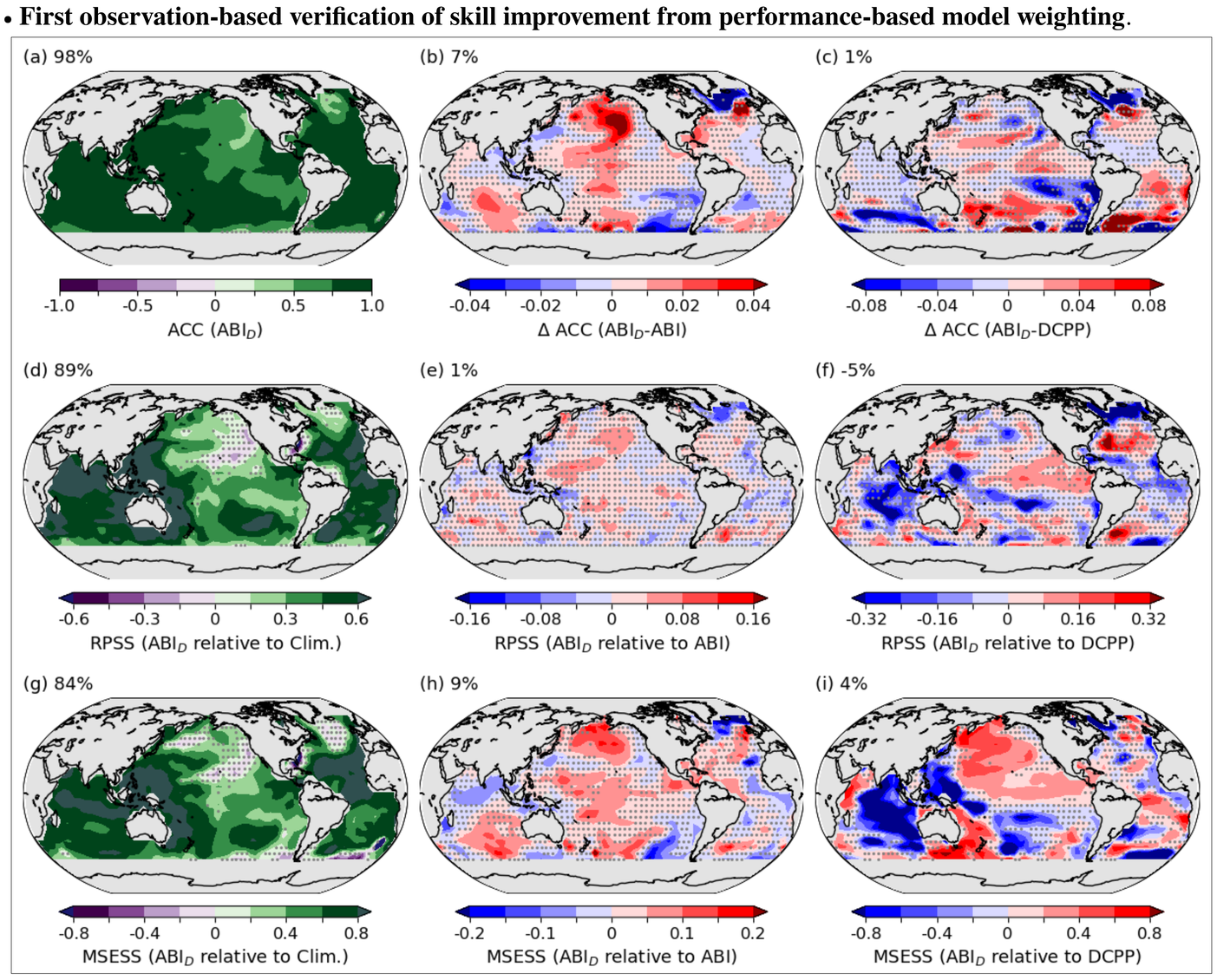

Figure 5: Skill in real-world hindcasts. (a) Anomaly correlation coefficient (ACC) of the ABI$_D$. Difference in ACC between the ABI$_D$ and (b) the ABI, and (c) the Decdal Climate Prediction Project (DCPP). Ranked probability skill score (RPSS) of the ABI$_D$ relative to (d) climatology, (e) ABI, and (f) DCPP as reference. Mean squared error skill score (MSESS) of the ABI$_D$ relative to (g) climatology, (h) the ABI, and (i) DCPP as reference. Skill is for SST average of hindcast years 1-10. Number of start dates is 86 (1930-2015) in (a,b,d,e,g,h) and 56 (1960-2015) in (c,f,i). 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.

EXPECT General Assembly 2026, Lisbon, Portugal

Funded by the European Union Horizon Europe, grant 101137656