Environmental Data Science (2026), 5: e21, 1–14
doi:10.1017/eds.2026.10044

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Interpretable machine learning for CMIP6 multi-model ensembles

Siyi Wu and Steve M. Easterbrook

University of Toronto, Canada
Corresponding author: Siyi Wu; Email: reyna.wu@mail.utoronto.ca

Received: 31 May 2026 UTC; **Revised:** 31 May 2026 UTC; **Accepted:** 04 June 2026 UTC

Keywords: CMIP6; clustering; interpretable machine learning; ensemble weighting; multi-model ensembles

Abstract

Multi-model ensembles are widely used in climate science, yet Coupled Model Intercomparison Project Phase 6 (CMIP6) models are neither independent nor equally skillful. We present a framework for interpretable ensemble learning that improves prediction while making model contributions explicit. Using monthly CMIP6 near-surface temperature and precipitation (1948–2014) against ERA5, we compare arithmetic multi-model ensemble (AMME) averaging with Ridge regression, random forest, and a continuous gating mixture that adaptively interpolates between them. On held-out test years (2006–2014), gating achieves the best temperature performance and improves precipitation distributional behavior relative to both AMME and individual learners. We trace these gains back to individual models and space: dominant-cluster and dominant-model maps show that gating reallocates trust in regionally coherent patterns rather than collapsing to a single best model, while entropy-based diagnostics quantify where trust is concentrated versus distributed. These analyses provide a practical pathway for transparent ensemble learning that supports both predictive skill and scientific accountability.

Impact Statement

Climate projections increasingly rely on combining many climate models, yet these models are not equally reliable or independent. This study shows how machine learning can improve multi-model ensemble predictions while keeping model contributions visible. Using CMIP6 temperature and precipitation simulations evaluated against ERA5, we compare arithmetic averaging with Ridge regression, random forest, and an adaptive gating approach. The gating model improves predictive skill, especially for temperature, while revealing where trust is concentrated in particular models, clusters, regions, or seasons. These diagnostics help researchers understand when data-driven weighting is useful rather than treating machine learning as a black box. The framework offers a transparent pathway for building more accountable climate ensemble tools for research and decision support in future climate assessments.

1. Introduction

Multi-model ensembles are a central tool in climate science, particularly for synthesizing uncertainty and producing robust climate projections. In large coordinated experiments such as CMIP, the dominant

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

practice is to combine models using simple arithmetic averaging, implicitly treating each model as an independent and equally plausible representation of the climate system. This "model democracy" approach remains widely used due to its simplicity, transparency, and long-standing institutional acceptance.

However, extensive work in climate science has shown that these assumptions rarely hold in practice. Climate models are neither independent nor equally skillful: many share code bases and parameterizations, and their performance varies substantially across variables, regions, and temporal scales (Sanderson et al., 2015; Knutti et al., 2017). As a result, equal weighting can lead to over-representation of closely related model families and suboptimal ensemble predictions. This has motivated a large body of research on performance-based and dependence-aware weighting schemes for climate ensembles (Knutti et al., 2017; Slater et al., 2017).

In parallel, recent studies have increasingly applied machine learning methods for climate ensemble post-processing, including linear regression, random forests (RFs), and neural networks (Dey et al., 2022; Pakdaman et al., 2022; Kumar and Dwarakish, 2025). These approaches often report improved predictive accuracy compared to unweighted means, especially for near-surface temperature and hydrological variables. However, most ML-based ensemble studies focus primarily on aggregate skill metrics and provide limited insight into how individual climate models contribute to the final prediction. This lack of interpretability makes it difficult to assess robustness, diagnose failure modes, or connect data-driven improvements back to physical understanding.

In this paper, we present an application-oriented framework for *interpretable ensemble learning* using CMIP6 simulations and ERA5 reanalysis. We track how machine learning methods reallocate trust across individual climate models and across coarse groupings of models defined by their climatological behavior. Our goal is to understand when machine learning is epistemically useful for climate ensemble post-processing, and which models drive those improvements.

We structure our analysis around three applied research questions:

- **RQ1 (Structure):** To what extent do simple climatological summaries reveal coarse groupings among CMIP6 models, and how useful are these groupings for interpreting ensemble learning behavior?
- **RQ2 (Skill):** How do machine learning ensemble methods compare to arithmetic averaging for predicting near-surface temperature and precipitation?
- **RQ3 (Interpretability):** Which climate models receive the highest effective weights under machine learning, and how do these weights vary across clusters, spatial regions, and seasons?

2. Data and experimental design

2.1. Climate model and observational data

We use monthly near-surface air temperature (tas) and precipitation (pr) from 25 CMIP6 global climate models, covering the historical period 1948–2014. The selected models span a wide range of modeling centers and include both Earth system models and coupled general circulation models. All model outputs are regridded to a common $1^\circ \times 1^\circ$ latitude–longitude grid to ensure spatial consistency across models.

As observational reference, we use the ERA5 reanalysis product, which provides globally complete and physically consistent estimates of atmospheric variables based on data assimilation of satellite and in situ observations. ERA5 is treated as the target for supervised learning and model evaluation. For both CMIP6 and ERA5, we compute global area-weighted means using cosine latitude weights.

2.2. Climatological feature construction

To characterize systematic differences among climate models, we compute each model's long-term climatological mean for temperature and precipitation. Specifically, for each CMIP6 model m, we define

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

$$\bar{x}_m = \frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{i,j} w_{ij} x_{m,t,i,j}}{\sum_{i,j} w_{ij}}, \tag{2.1}$$

where $x_{m,t,i,j}$ denotes the variable value at time t and grid cell (i,j), and $w_{ij} = \cos(\phi_{ij})$ are latitude-based area weights. These climatological summaries provide a low-dimensional representation of each model's large-scale behavior and are used as features for clustering.

We intentionally adopt a low-dimensional representation based on long-term global climatological means. Our goal is to construct a coarse, interpretable partition that can be directly related to how ensemble learning redistributes weights. A scalar global mean is the simplest such feature and the most transparent for tracing learned weights. This choice reflects a trade-off between representational richness and interpretability: while higher-dimensional features (e.g., regional or decadal statistics) may capture more detailed structure, they also make it more difficult to attribute learned weights back to physically meaningful differences among models. We acknowledge this limitation explicitly; exploring richer feature spaces is an important direction for future work.

2.3. Train–validation–test splitting

All supervised learning experiments follow a time-ordered split to avoid temporal leakage. The full time series is divided into a training period (1948–1998), a validation period (1999–2005), and a test period (2006–2014). Machine learning models are trained on the training set, hyperparameters are selected using the validation set, and all reported performance metrics are computed on the held-out test period.

2.4. Ensemble learning methods

We compare four ensemble post-processing strategies. As a baseline, the arithmetic multi-model ensemble (AMME) prediction is computed as the simple average across all CMIP6 models:

$$\hat{y}_t^{\mathrm{AMME}} = \frac{1}{M}\sum_{m=1}^{M} x_{m,t}, \tag{2.2}$$

where M denotes the number of models and $x_{m,t}$ the prediction from model m at time t.

We then fit a Ridge regression model to learn a global linear weighting of models,

$$\hat{y}_t^{\mathrm{Ridge}} = \sum_{m=1}^{M} w_m x_{m,t} + b, \tag{2.3}$$

where weights w_m are estimated by minimizing squared error subject to an ℓ_2 regularization penalty. Hyperparameter α is selected from $\{0.01, 0.1, 1.0, 10.0, 100.0\}$ on the validation set. This formulation yields an interpretable mapping between individual climate models and their contributions to the ensemble prediction. The ℓ_2 penalty distributes weight broadly across models, preventing any single model from dominating; a sparser alternative such as LASSO (ℓ_1) would perform automatic model selection and is a natural direction for future work.

To capture nonlinear interactions among models, we also apply an RF regressor using model outputs as input features. We use $n_{\mathrm{estimators}} = 600$ trees with min_samples_leaf = 2. Variable importance scores are computed as the mean decrease in impurity (MDI) averaged over all trees and normalized to sum to 1; they quantify the relative contribution of each CMIP6 model to the RF prediction.

Finally, we implement a continuous gating mixture model that learns a convex combination of Ridge and RF predictions. The gated prediction is

$$\hat{y}_{t,i,j}^{\mathrm{Gate}} = g_{t,i,j}\hat{y}_{t,i,j}^{\mathrm{Ridge}} + (1 - g_{t,i,j})\hat{y}_{t,i,j}^{\mathrm{RF}}, \tag{2.4}$$

where subscripts (i,j) denote grid cell and t denotes time (global mean analyses use scalar g_t without spatial indices). For each validation time step, the optimal gate value is

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

$$g_t^* = \mathrm{clip}\left(\frac{\left(y_t - \hat{y}_t^{\mathrm{RF}}\right)\left(\hat{y}_t^{\mathrm{Ridge}} - \hat{y}_t^{\mathrm{RF}}\right)}{\left(\hat{y}_t^{\mathrm{Ridge}} - \hat{y}_t^{\mathrm{RF}}\right)^2 + \varepsilon}, 0, 1\right), \tag{2.5}$$

where $\varepsilon = 10^{-12}$ ensures numerical stability. A second Ridge regression model is then fitted on the validation set to predict g_t^* from the standardized CMIP6 inputs $\mathbf{x}_t$. At test time, the gate g_t is evaluated by applying this fitted model to the (standardized) test inputs, so no test-period information is used in the gate.

The effective per-model gating weight is defined as

$$w_m^{\mathrm{Gate}} = \overline{g}\,\tilde{w}_m^{\mathrm{Ridge}} + (1 - \overline{g})\, I_m^{\mathrm{RF}}, \tag{2.6}$$

where $\overline{g}$ is the time-averaged gate value, $\tilde{w}_m^{\mathrm{Ridge}} = |w_m| / \sum_{m'} |w_{m'}|$ is the normalized Ridge weight, and I_m^{RF} is the RF importance for model m. These effective weights are used in the spatial analyses of Section 5.

2.5. Evaluation metrics

Model performance is evaluated using multiple complementary metrics, including root mean square error (RMSE), mean absolute error (MAE), Pearson correlation, Nash–Sutcliffe efficiency (NSE), variance ratio between predictions and observations, and predictive skill relative to the arithmetic ensemble:

$$\mathrm{Skill} = 1 - \frac{\mathrm{RMSE}_{\mathrm{method}}}{\mathrm{RMSE}_{\mathrm{AMME}}}. \tag{2.7}$$

Positive skill indicates improvement over AMME.

For precipitation, we additionally report distribution-aware metrics. *Wet-day frequency RMSE* is the RMSE of the fraction of time steps exceeding a threshold of 0.1 mm day^{-1}, computed between each method's monthly predictions and ERA5 over the test period. *Upper-tail RMSE* is the RMSE restricted to time steps where ERA5 exceeds its 90th percentile of the observed precipitation distribution, providing a measure of extreme-event representation.

Trust entropy is defined as

$$H_{i,j} = -\sum_{m=1}^{M} p_{m,i,j} \log\left(p_{m,i,j} + \varepsilon\right),\ p_{m,i,j} = \frac{|w_{m,i,j}|}{\sum_{m'} |w_{m',i,j}| + \varepsilon}, \tag{2.8}$$

where $w_{m,i,j}$ is the effective weight of model m at grid cell (i,j). The effective ensemble size is $N_{\mathrm{eff},i,j} = \exp\left(H_{i,j}\right)$, ranging from 1 (full concentration on a single model) to M (uniform trust across all models).

3. Clustering of CMIP6 models

We cluster CMIP6 models using a low-dimensional characterization of large-scale climatology: each model is represented by its global area-weighted mean temperature and precipitation over 1948–2014. This representation is intended as a descriptive summary of coarse inter-model differences rather than a definitive characterization of structural similarity among models. We apply k-means clustering after standardizing both variables and select k by comparing silhouette scores (Table 1). Among $k \in \{2,3,4,5\}$, $k = 4$ yields the best joint silhouette score across both temperature (0.649) and precipitation (0.569); $k = 5$ improves temperature marginally but reduces the precipitation score below $k = 4$, so $k = 4$ is preferred on joint optimality.

We note that alternative partitioning strategies – such as fixed-interval bins of mean bias relative to ERA5, or quantile-based groupings – would be equally motivated. K-means has the advantage of being data-adaptive (bin boundaries are determined by the data rather than fixed a priori), but the resulting groups are approximate and should not be interpreted as sharply separated categories.

Figure 1 shows the clustering results for both variables, with models ordered by their climatological temperature value so that cross-variable cluster shifts are immediately visible. Temperature exhibits

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

Table 1. *Average silhouette scores for k-means clustering of CMIP6 models based on climatological mean temperature and precipitation*

Number of clusters (k)	Temperature silhouette	Precipitation silhouette
2	0.621	0.616
3	0.607	0.507
4	**0.649**	**0.569**
5	0.643	0.553

Note. Higher values indicate better cluster separation. k = 4 is selected on joint optimality across both variables.

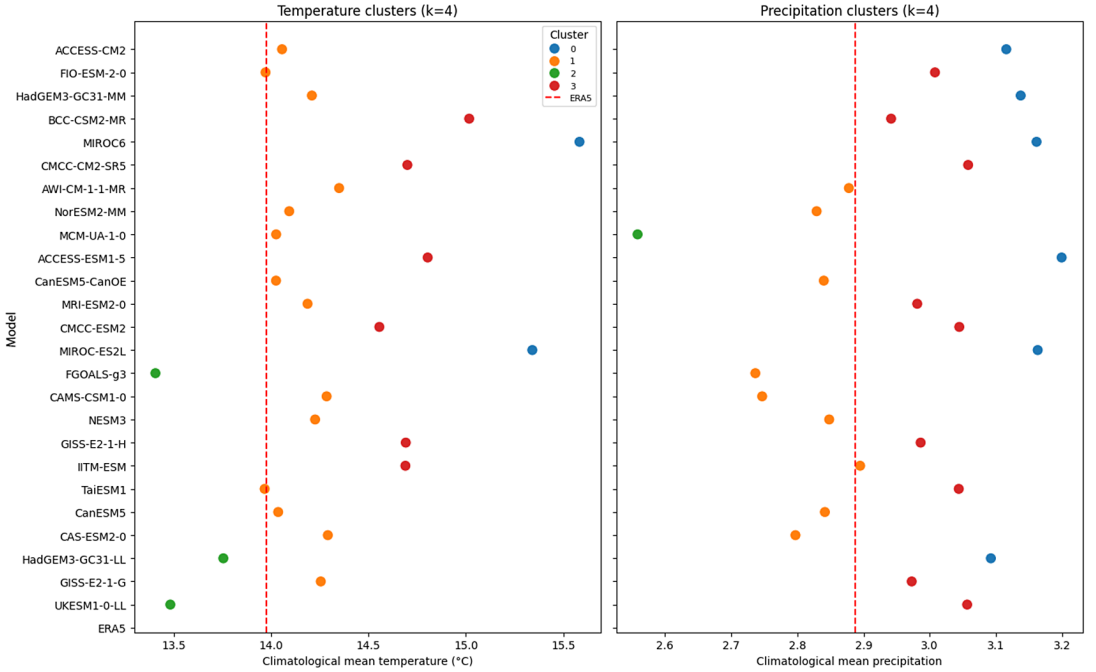


Figure 1. *Clustering of CMIP6 models based on climatological mean temperature (left) and precipitation (right), using k-means with k = 4. Models are ordered by their climatological temperature value on the shared y-axis, so that color changes at the same y-position (between the two panels) indicate that a model's cluster assignment differs between the two variables. The red dashed line indicates the ERA5 climatological mean. Colors indicate cluster membership (consistent with all subsequent figures).*

moderate separation (silhouette 0.649), while precipitation displays greater overlap across clusters (silhouette 0.569). This difference likely reflects the higher variability and intermittency of precipitation, but we emphasize that the resulting clusters should be interpreted as approximate groupings rather than sharply separated categories.

Importantly, cluster membership depends on the target variable: some models remain grouped across variables, whereas others shift between temperature- and precipitation-based clusters, as is visible in Figure 1 where color changes at the same y-position indicate a cluster shift between the two panels. These patterns highlight that even simple climatological descriptors can capture meaningful differences in model behavior, though they do not directly imply shared model structure or common development lineage.

Preliminary experiments clustering models jointly on standardized temperature and precipitation global means produced broadly similar dominant groupings but with less stable cluster membership

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

across random seeds, suggesting that clustering outcomes are sensitive to feature choice. Richer representations – such as regional averages, multi-decadal statistics, or joint temperature–precipitation features – would likely yield different groupings and are an important direction for future work. In this work, we therefore treat clustering primarily as an exploratory tool to organize model diversity and to provide an interpretive lens for the weight analyses of Section 5.

4. Global prediction performance

We evaluate all ensemble methods on held-out test years (2006–2014) using multiple complementary metrics. Tables 2 and 3 report global results for temperature and precipitation, respectively.

4.1. Temperature

All learning-based methods outperform AMME for global mean temperature (Table 2). Ridge reduces RMSE by more than half relative to AMME, and RF also yields substantial improvement. The gating mixture achieves the best overall performance, with the lowest RMSE (0.196) and highest NSE (0.980). Correlations with ERA5 remain very high across methods, indicating that improvements reflect better tracking of temporal variability rather than simple variance rescaling.

To confirm that these rankings are not artifacts of a single training run, we repeated the experiment with $n = 10$ different random seeds for the RF (Ridge is deterministic given a fixed training set). Table 4 reports mean RMSE and standard deviation across seeds. The standard deviation is near zero for all methods (Ridge: 0.000, RF: 0.002, Gating: 0.001), confirming that the ranking Gating > Ridge > RF > AMME is stable and not attributable to random variation in tree construction.

4.2. Precipitation

For precipitation, we report distribution-aware skill using RMSE, wet-day frequency RMSE, and upper-tail RMSE (Table 3). All learning-based approaches improve over AMME across these metrics. Ridge

Table 2. *Global prediction performance for near-surface temperature (tas) on the test period 2006–2014*

Method	RMSE	MAE	Corr	NSE	VarRatio	Skill vs. AMME
AMME	0.493	0.475	**0.996**	0.873	1.024	–
Ridge	0.209	0.184	**0.996**	0.977	**0.992**	0.576
RF	0.225	0.195	0.992	0.974	0.953	0.543
Gating	**0.196**	**0.169**	**0.996**	**0.980**	0.967	**0.604**

Note. Bold indicates the best value in each column. For VarRatio, bold indicates the value closest to 1.0 (perfect variance calibration). Best Corr is shared by AMME, Ridge, and Gating (all 0.996).

Table 3. *Global precipitation prediction performance on the test period 2006–2014*

Method	RMSE	MAE	WetFreq RMSE	Tail RMSE	Skill vs. AMME
AMME	0.00444	0.00381	0.3168	0.00684	–
Ridge	0.00243	0.00198	0.2165	0.00358	0.452
RF	0.00278	0.00229	0.2274	0.00428	0.374
Gating	**0.00218**	**0.00176**	**0.2093**	**0.00319**	**0.509**

Note. Wet-day frequency RMSE measures the error in reproducing the fraction of time steps exceeding 0.1 mm day^{-1}; upper-tail RMSE is restricted to time steps where ERA5 exceeds its 90th percentile. Skill vs. AMME $= 1 - \mathrm{RMSE}_{\mathrm{method}}/\mathrm{RMSE}_{\mathrm{AMME}}$. Bold indicates the best value in each column.

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

Table 4. *Stability of global temperature RMSE across $n = 10$ random seeds*

	Ridge	RF	Gating
Mean RMSE	0.209	0.227	0.196
Std RMSE	0.000	0.002	0.001

Note. Ridge is deterministic (std = 0). The near-zero standard deviations confirm that the method ranking is not attributable to random variation.

provides strong gains relative to AMME, while RF improves less consistently. The gating mixture yields the best overall performance globally, improving not only mean error but also wet-day frequency and tail behavior. Compared to temperature, precipitation gains are smaller and more sensitive to method choice, consistent with higher intermittency and noise in the precipitation field.

Figure 2 provides a qualitative illustration of the test-period performance. Panel (a) shows all 25 CMIP6 model time series color coded by Ridge regression weight: the most-trusted models (red) track ERA5 closely, while less-trusted models (pale/blue) deviate more. Panel (b) uses the same time series but codes color by RF importance, revealing which models the nonlinear learner relies on most. Panel (c) compares predicted vs. observed global mean temperature for all four methods on the test period (2006–2014): all learning-based methods closely track ERA5, while AMME shows a systematic warm bias reflecting the multi-model mean offset.

5. Interpretable model weights & gating behavior

While ensemble learning methods improve predictive skill, their scientific value ultimately depends on whether they provide interpretable insight into the role of individual climate models. We therefore analyze the effective weights assigned to each CMIP6 model under Ridge regression and the gating mixture model, and examine how these weights vary across clusters, spatial regions, and seasons.

5.1. Global ridge weights and comparison with RF

Table 5 lists the ten models receiving the highest global Ridge weights for temperature, together with their RF importance scores. These models contribute heterogeneously to the ensemble prediction: the top-weighted models receive approximately 1.5 × the uniform AMME weight ($1/25 = 0.040$), while the lowest-weighted models receive correspondingly less. The degree of concentration is controlled by the ℓ_2 regularization strength: larger α pushes weights toward uniformity, while smaller α allows greater differentiation.

Notably, the highest-weighted models are concentrated within Clusters 1 and 3 for Ridge, suggesting that Ridge regression preferentially amplifies specific model families rather than uniformly redistributing trust across the ensemble.

Comparing Ridge weights with RF importances (Table 5 and Figure 3) reveals partial but imperfect agreement (Pearson $r = 0.52$). Six models receive substantially higher RF importance than Ridge weight: GISS-E2-1-H (RF: 0.236 vs. Ridge: 0.050), AWI-CM-1-1-MR (0.158 vs. 0.050), CanESM5 (0.143 vs. 0.046), MRI-ESM2–0 (0.123 vs. 0.050), FIO-ESM-2-0 (0.108 vs. 0.048), and FGOALS-g3 (0.106 vs. 0.044). These models appear to carry nonlinear information that the RF can exploit but that Ridge, as a globally linear model, cannot fully capture. Conversely, many models receive higher Ridge weight than RF importance, reflecting the regularized distribution of weights that Ridge enforces.

Figure 3 illustrates this comparison across all 25 models. Panel (a) shows normalized Ridge weights and RF importances as a grouped bar chart, with the uniform AMME baseline shown as a dashed line. Panel (b) shows a scatter plot of the two quantities; models above the identity line are RF-preferred, while models below are Ridge-preferred. The six strongly RF-preferred models cluster in the upper-left region. The partial agreement ($r = 0.52$) is interpreted in Section 5.2: the gating exploits both sets of signals by adaptively blending the two learners.

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

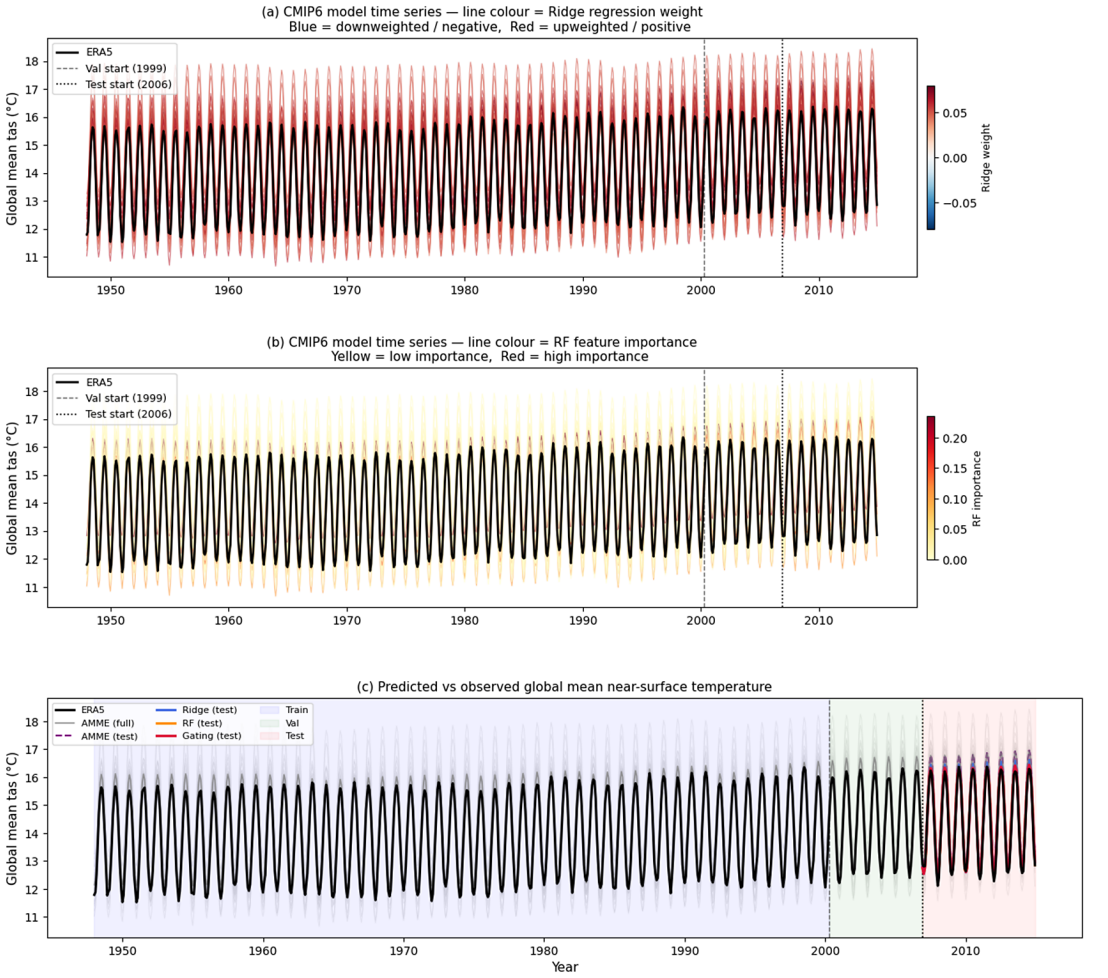


Figure 2. *Global mean near-surface temperature time series and test-period comparison. (a) All 25 CMIP6 model time series color coded by normalized Ridge regression weight: red indicates high positive weight, blue indicates negative or near-zero weight. ERA5 is shown in black. Training/validation/test periods are delineated by vertical dashed lines. (b) Same time series color coded by RF variable importance (yellow = low, red = high). (c) Predicted vs. observed global mean temperature for AMME (dashed pink) and Gating (solid red) on the test period 2006–2014. Ridge and Gating closely track ERA5, while AMME exhibits a systematic warm bias.*

5.2. Spatial structure of gating and model dominance

To examine how the gating ensemble redistributes trust across space, we analyze the dominant-cluster and dominant-model at each grid cell, based on the mean per-model gating weight (i.e., the average absolute weight per model within each cluster, rather than the sum, to avoid bias toward whichever cluster has the most members).

For spatial reference, we define three large-scale regions: the Tropics (30 ° S–30 ° N), Northern Hemisphere mid-latitudes (30 ° N–60 ° N), and the Arctic (60 ° N–90 ° N). The mean gating skill improvement over AMME is +0.418 in the Tropics, +0.148 in NH mid-latitudes, +0.319 in the Southern Hemisphere mid-latitudes, +0.298 in the Arctic, and +0.294 in the Antarctic.

The left panel of Figure 4 shows the dominant-model cluster per grid cell. Clear large-scale spatial organization emerges, with distinct cluster regimes in the tropics, mid-latitudes, and high-latitude oceans.

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

Table 5. *Top ten CMIP6 models by global Ridge weight for temperature, together with their cluster assignment (consistent with the fixed-seed clustering reported in Section 3)*

Model	Ridge Weight	Cluster	RF Importance
CMCC-CM2-SR5	0.0613	3	0.024
GISS-E2-1-H	0.0575	3	0.236
CMCC-ESM2	0.0573	3	0.005
CAMS-CSM1-0	0.0522	1	0.000
AWI-CM-1-1-MR	0.0502	1	0.158
MRI-ESM2-0	0.0501	1	0.123
GISS-E2-1-G	0.0495	1	0.014
FIO-ESM-2-0	0.0464	1	0.108
CanESM5	0.0461	1	0.143
CanESM5-CanOE	0.0440	1	0.007

Note. The rightmost column shows the corresponding RF importance score for comparison.

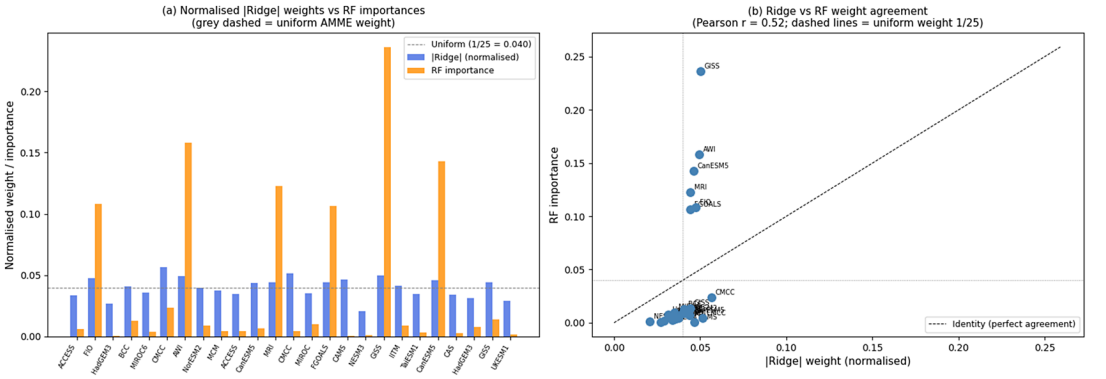


Figure 3. *(a) Normalized Ridge regression weights (blue) and RF variable importances (orange) for all 25 CMIP6 models, sorted by Ridge weight. The dashed gray line shows the uniform AMME weight (1/25). (b) Scatter of normalized Ridge weight vs. RF importance per model (Pearson $r = 0.52$). Models above the diagonal are more trusted by RF than Ridge; the six models with RF importance $> 2\times$ Ridge weight are labeled. The partial agreement indicates that the two learners identify overlapping but non-identical sets of informative models.*

Cluster 0 (MIROC models) dominates parts of the tropical and subtropical oceans; Cluster 3 (CMCC, GISS-H, ACCESS-ESM1–5 group) is preferred over continental interiors and parts of the Southern Ocean; Cluster 1 (the large 14-member group) dominates high-latitude regions. These spatial patterns are broadly consistent with known dynamical differences among climate model families: for example, the MIROC models are known for distinctive SST biases in the tropical Pacific, while the CMCC and GISS families differ in aerosol forcing and ocean heat uptake representations (Sanderson et al., 2015; Knutti et al., 2017). The right panel of Figure 4 shows the dominant individual CMIP6 model per grid cell. Rather than a single model dominating globally, the gating ensemble exhibits strong spatial diversity, with different models preferred in different regions.

To understand how gating relates to its two components, Figure 5 shows dominant-cluster maps and trust entropy maps for Ridge, RF, and Gating side by side. Ridge produces a spatially smooth, multi-cluster pattern (mean $N_{\text{eff}} = 21.1$), while RF is more concentrated (mean $N_{\text{eff}} = 13.4$), reflecting the implicit feature selection of tree-based methods. Gating achieves an intermediate effective ensemble size

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

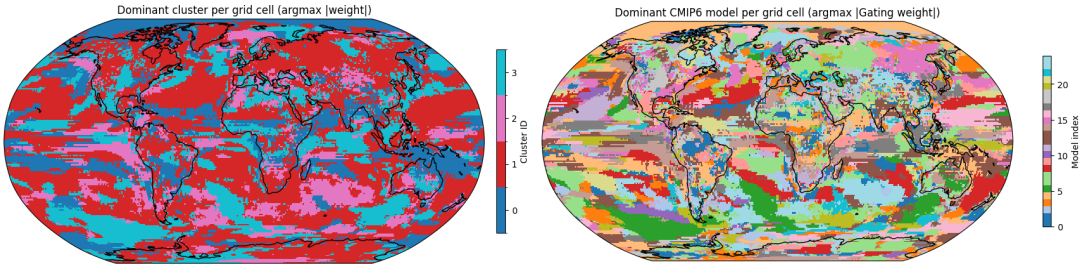


Figure 4. *(Left) Dominant CMIP6 model cluster per grid cell, defined by the cluster whose models receive the highest mean per-model gating weight. (Right) Dominant individual CMIP6 model per grid cell, defined by the model with the maximum average gating weight. Results show strong spatial structure in ensemble trust, with different clusters and models preferred across distinct dynamical regimes. Cluster IDs: 0 = MIROC family (2 models), 1 = large mid-bias group (14 models), 2 = HadGEM–UKESM family (3 models), 3 = CMCC–GISS-H group (6 models).*

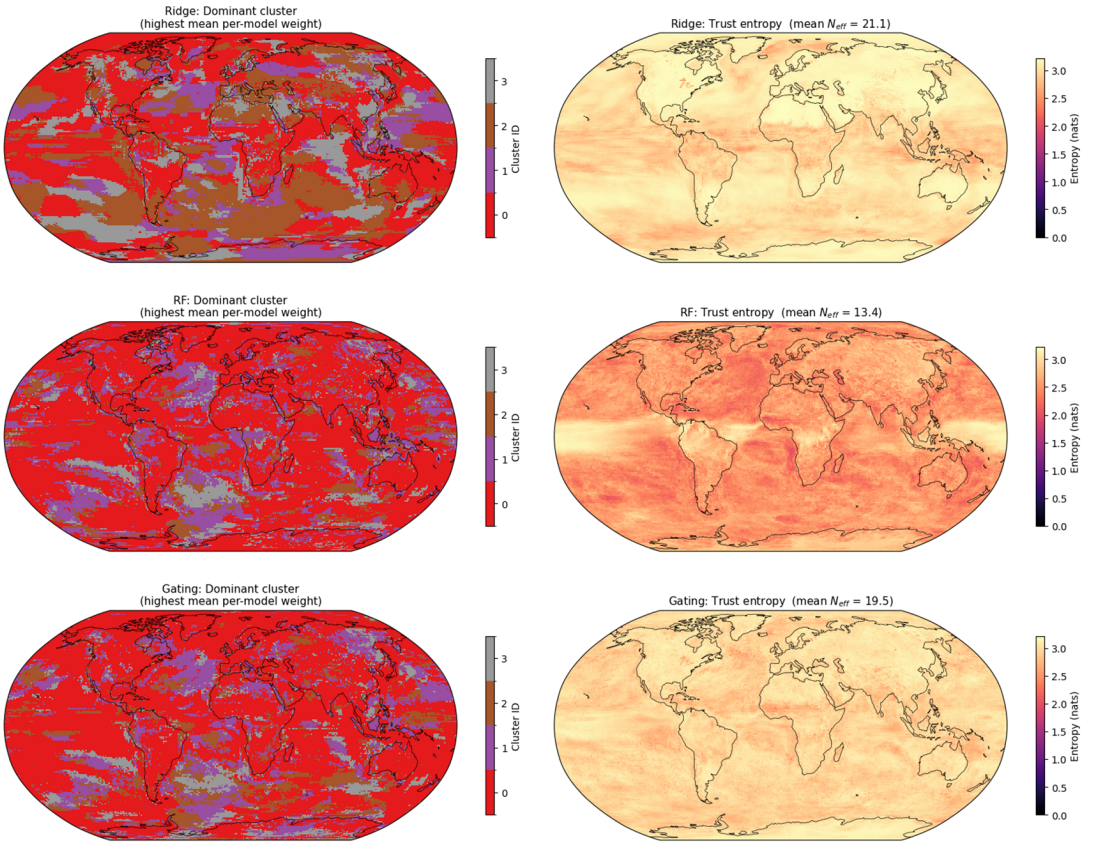


Figure 5. *Dominant cluster (left column) and trust entropy (right column) for Ridge (top), RF (middle), and Gating (bottom). Ridge maintains high entropy (mean $N_{\text{eff}} = 21.1$) with a spatially smooth multi-cluster pattern. RF is more concentrated (mean $N_{\text{eff}} = 13.4$), with stronger Cluster-0 (MIROC) dominance in the tropics. Gating achieves an intermediate $N_{\text{eff}} = 19.5$, selectively concentrating trust in regions where RF's nonlinear predictions are locally superior.*

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

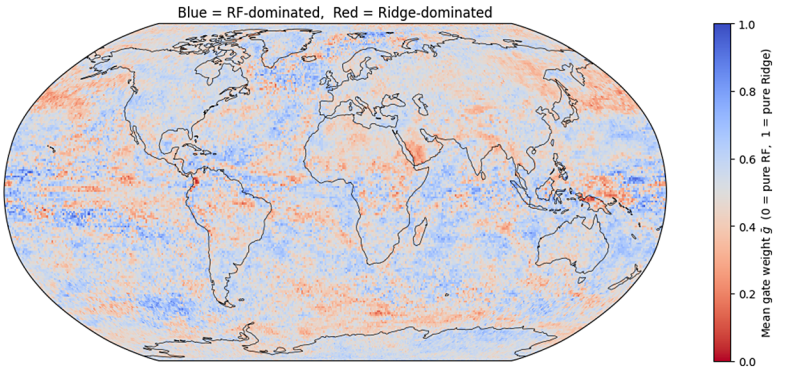


Figure 6. *Spatial distribution of the time-mean gate value $\overline{g}$ (per pixel). Red ($\overline{g}$ near 1) indicates Ridge-dominated pixels; blue ($\overline{g}$ near 0) indicates RF-dominated pixels. Overall, 55% of pixels are Ridge-dominated and 45% are RF-dominated. RF dominance is concentrated in tropical oceans and the Southern Ocean, which also exhibit the strongest skill improvement relative to AMME (Figure 7).*

($N_{\text{eff}} = 19.5$), broadly following the Ridge spatial pattern but selectively concentrating trust in regions where RF's nonlinear predictions are locally superior – consistent with the gating mechanism blending both learners. The spatial distribution of the per-pixel mean gate value $\overline{g}$ (Figure 6) confirms this: 55% of pixels are Ridge-dominated ($\overline{g} > 0.5$) and 45% are RF-dominated ($\overline{g} < 0.5$), with RF dominance concentrated in tropical and Southern Ocean regions that also exhibit the strongest skill improvement.

Figure 7 shows the spatial distribution of skill improvement of the gating ensemble relative to AMME for near-surface temperature. Large contiguous regions exhibit substantial positive skill, particularly over tropical oceans, the Southern Ocean, and parts of the Indian Ocean basin. These regions correspond to areas of strong model disagreement and known systematic biases in coupled climate models, suggesting that gating is most effective where ensemble diversity is high and model errors are structured rather than random. Conversely, regions with weak or near-zero skill gains are primarily located over continental interiors and high-latitude land areas, where internal variability dominates and cross-model differences are less informative. This spatial pattern indicates that gating does not uniformly improve predictions everywhere but selectively enhances skill in regions where ensemble learning is epistemically meaningful.

Compared to traditional performance-based weighting schemes – such as scalar skill-based weighting (Knutti et al., 2017) or the ClimWIP approach (Sanderson et al., 2015), which assign globally uniform scalar weights – the gating mechanism operates at the pixel level and is time-varying, allowing trust to be concentrated only where and when a sub-model is locally superior. The ℓ_2 regularization in Ridge explains why that component maintains high entropy (many models contribute); the tree-based selection of RF explains its lower entropy; and gating mediates between these extremes.

5.3. Trust dispersion and effective ensemble size

To quantify how concentrated or distributed model trust is under the gating ensemble, we compute the entropy of the gating weights at each grid cell (Section 2.5). Higher entropy indicates more evenly distributed trust across models, while lower entropy indicates strong concentration on a small number of models.

Figure 8 shows the spatial distribution of trust entropy for the gating ensemble. Most regions exhibit relatively high entropy, corresponding to a mean effective ensemble size of $N_{\text{eff}} = 19.5$ models. For comparison, Ridge yields $N_{\text{eff}} = 21.1$ and RF yields $N_{\text{eff}} = 13.4$ (see also Figure 5). This indicates that, despite spatially varying dominance patterns, gating generally maintains a broad ensemble rather than collapsing to a few preferred models.

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

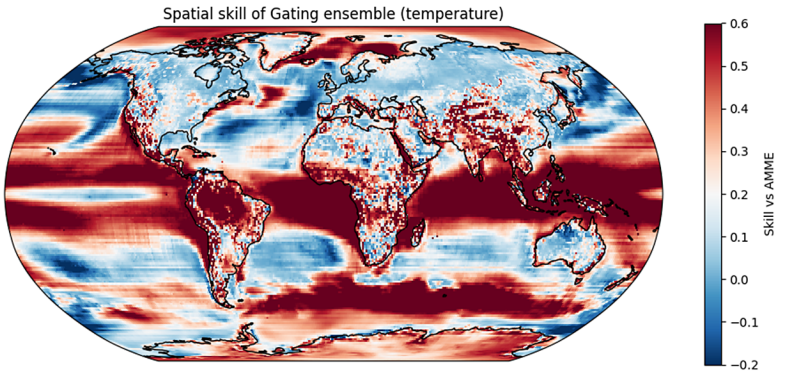


Figure 7. *Spatial distribution of skill improvement of the gating ensemble relative to AMME for near-surface temperature (test period 2006–2014). The diverging colormap is centered at zero (white), with red indicating improvement and blue indicating degradation relative to AMME. Positive values indicate regions where gating reduces RMSE compared to the arithmetic ensemble. Strong improvements occur over tropical oceans and parts of the Southern Hemisphere; small degradations occur over some continental interiors and high-latitude land areas.*

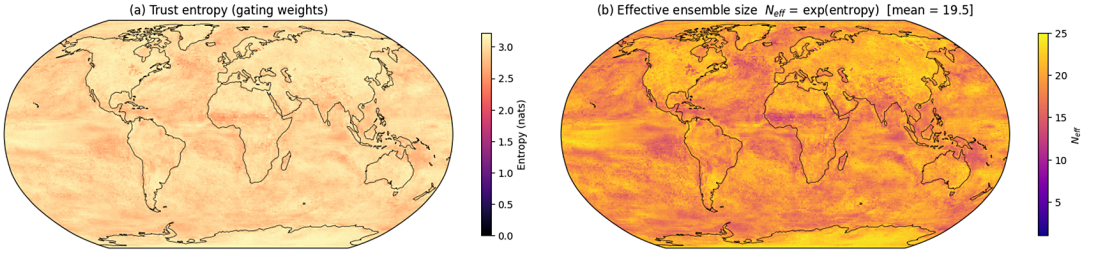


Figure 8. *Spatial distribution of trust dispersion under the Gating ensemble, measured as (a) trust entropy H and (b) effective ensemble size $N_{\text{eff}} = \exp(H)$ [mean $= 19.5$ models]. Higher values indicate more evenly distributed trust (larger effective ensemble size), while lower values indicate strong concentration on a small subset of models. For comparison, Ridge yields mean $N_{\text{eff}} = 21.1$ and RF yields mean $N_{\text{eff}} = 13.4$; see Figure 5 for side-by-side entropy maps of all three methods.*

Regions with lower entropy appear primarily in parts of the tropical oceans and the Southern Ocean, where a small subset of models consistently outperforms others. These regions also correspond to areas of particularly strong skill improvement under gating (Figure 7), suggesting that localized trust concentration is associated with genuine predictive advantage.

Taken together, these results demonstrate that gating produces a structured redistribution of trust: locally selective where strong signals exist, but globally conservative in preserving ensemble diversity. This behavior supports the interpretation of gating as a principled trust reallocation mechanism rather than an aggressive model selection procedure.

5.4. Temporal and seasonal behavior of the gate

To assess the temporal consistency of the gating weights, Figure 9 shows the gate value g_t over the test period 2006–2014 together with a 12-month rolling mean and a per-season breakdown. The gate is Ridge-dominated for 92% of monthly time steps (global mean $\overline{g} = 0.685$) but exhibits meaningful seasonal modulation. MAM (spring) has the highest mean gate value ($\overline{g}_{\text{MAM}} = 0.873$, almost entirely Ridge-

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

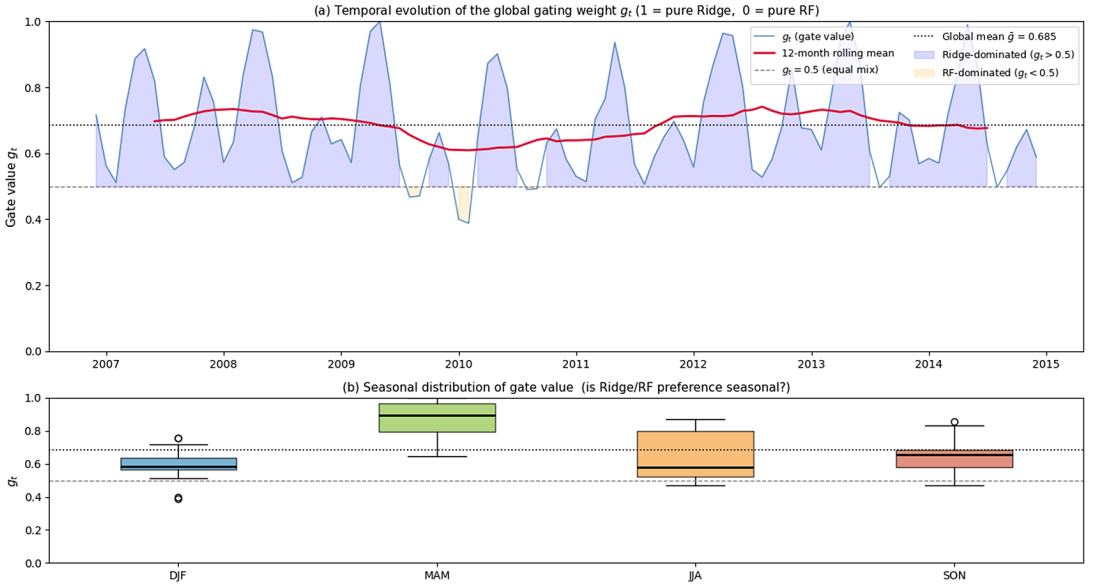


Figure 9. *(a) Temporal evolution of the global gating weight g_t over the test period 2006–2014. Blue shading indicates Ridge-dominated months ($g_t > 0.5$); yellow shading indicates RF-dominated months ($g_t < 0.5$). The red curve shows a 12-month rolling mean; the dotted horizontal line shows the global mean $\overline{g} = 0.685$. (b) Seasonal distribution of g_t: MAM is strongly Ridge-dominated (mean 0.873), while DJF shows the greatest RF contribution (mean 0.592). Boxplots show the interquartile range and outliers across individual months within each season.*

dominated), while DJF (winter) has the lowest ($\overline{g}_{\mathrm{DJF}} = 0.592$), with JJA and SON at intermediate values (0.638 and 0.639, respectively). This pattern suggests that the global mean relationship between CMIP6 models and ERA5 is more linear in spring – when ensemble spread may be dominated by systematic rather than nonlinear biases – and that the RF nonlinearity contributes more in winter. We note that this temporal gate operates on the global mean signal; spatial variation of $\overline{g}$ is shown in Figure 6. Seasonal effects on the spatial gating weights are a natural direction for future work.

6. Conclusion

This paper presented an interpretable ensemble learning framework that improves prediction while making the contributions of individual CMIP6 models explicit. We first clarified that the clustering analysis is an exploratory organizational tool rather than a definitive characterization of model structural similarity: *k*-means on global climatological means provides a coarse, interpretable partition that can be directly related to how ensemble learning redistributes weights but does not recover code-lineage or parameterization relationships. We showed that CMIP6 exhibits structured diversity: temperature supports moderate clustering and precipitation displays greater overlap. Building on this, we compared four ensemble strategies against ERA5 using a time-ordered split. For near-surface temperature, all learning-based methods substantially outperformed AMME, with the gating mixture achieving the best performance; stability analysis over ten random seeds confirms these rankings are robust. For precipitation, distribution-aware evaluation showed that learning-based ensembles improve not only mean error but also wet-day frequency and tail behavior, with gating again yielding the strongest global results.

Crucially, improvements arise from structured reallocation of trust rather than global selection of a single model. Spatial analyses reveal that gating concentrates weights only where doing so provides clear benefit, while maintaining a broad effective ensemble ($N_{\mathrm{eff}} \approx 19.5$) elsewhere. A comparison of Ridge,

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press

RF, and gating spatial weight patterns shows that RF is more concentrated ($N_{\text{eff}} = 13.4$) while Ridge is more distributed ($N_{\text{eff}} = 21.1$); gating mediates between these extremes, exploiting both the linear regularization of Ridge and the nonlinear feature selection of RF. Temporal analysis of the gate shows meaningful seasonal variation, with Ridge-dominated behavior in spring and greater RF contribution in winter, though causal attribution of this pattern to specific physical processes remains an open question.

These results suggest a practical middle ground between model democracy and aggressive model selection: ensemble learning can improve skill while remaining transparent and auditable. Future work should extend this approach to richer model descriptors (e.g., regional statistics, multi-decadal trends), uncertainty-aware objectives, sparse weighting schemes (e.g., LASSO), joint multivariate learning across temperature and precipitation, and additional observational products, and should assess whether seasonal and regional patterns in gating weights can be attributed to specific physical mechanisms in the underlying models.

Acknowledgments. We acknowledge the World Climate Research Programme (WCRP), which, through its Coupled Model Intercomparison Project (CMIP), coordinated the CMIP6 experiments used in this study. We thank the climate modeling groups for producing and making their model outputs publicly available, the Earth System Grid Federation (ESGF) for archiving and distributing CMIP6 data, and the European Centre for Medium-Range Weather Forecasts (ECMWF) for providing the ERA5 reanalysis dataset.

Author contribution. Conceptualization: S.W. Methodology: S.W. Software: S.E. Formal analysis: S.W. Data curation: S.W. Visualization: S.W. Writing – original draft preparation: S.W. Writing – review and editing: S.E., S.W. Supervision: S.E. All authors discussed the results and approved the final manuscript. All authors approved the final submitted draft.

Competing interests. The author(s) declare none.

Data availability statement. CMIP6 model outputs are publicly available through the Earth System Grid Federation (ESGF) data portals (https://esgf-node.llnl.gov/). ERA5 reanalysis data are available from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/). The analysis code and processed scripts used to reproduce all results and figures in this study are openly available on GitHub and archived on Zenodo at https://doi.org/10.5281/zenodo.19433526 under an MIT license.

Ethics statement. This study uses publicly available climate model simulations and reanalysis products and does not involve human participants, animals, or personal data. The research complies with all applicable institutional and international guidelines for the use of publicly available scientific data.

Funding statement. This work was supported by the Canadian Network for Research and Innovation in Machining Technology and the Natural Sciences and Engineering Research Council of Canada (NSERC), grant number RGPIN-202.

Provenance statement. This article is part of the Climate Informatics 2026 proceedings and was accepted in Environmental Data Science on the basis of the Climate Informatics peer review process.

References

Dey A, **Sahoo DP**, **Kumar R and Remesan R** (2022) A multimodel ensemble machine learning approach forCMIP6 climate model projections in an Indian River basin. *International Journal of Climatology 42*(16), 9215–9236. https://doi.org/10.1002/joc.7813.

Knutti R, **Sedláček J**, **Sanderson BM**, **Lorenz R**, **Fischer EM and Eyring V** (2017) A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters 44*(4), 1909–1918. https://doi.org/10.1002/2016GL072012.

Kumar GP and Dwarakish G (2025) Machine learning-based ensemble of global climate models and trend analysis for projecting extreme precipitation indices under future climate scenarios. *Environmental Monitoring and Assessment 197*(9), 1009. https://doi.org/10.1007/s10661-025-14469-6.

Pakdaman M, **Babaeian I and Bouwer LM** (2022) Improved monthly and seasonal multi-model ensemble precipitation forecasts in southwest asia using machine learning algorithms. *Water 14*(17), 2632. https://doi.org/10.3390/w14172632.

Sanderson BM, **Knutti R and Caldwell P** (2015) A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate 28*(13), 5171–5194. https://doi.org/10.1175/JCLI-D-14-00362.1.

Slater LJ, **Villarini G and Bradley AA** (2017) Weighting of NMME temperature and precipitation forecasts across Europe. *Journal of Hydrology 552*, 646–659. https://doi.org/10.1016/j.jhydrol.2017.07.029.

Cite this article: Wu S and Easterbrook SM (2026). Interpretable machine learning for CMIP6 multi-model ensembles. *Environmental Data Science*, 5: e21. doi:10.1017/eds.2026.10044

https://doi.org/10.1017/eds.2026.10044 Published online by Cambridge University Press