SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Hybrid Machine Learning Framework for Ordinal Classification Using Ordinal Logistic Regression and Calibrated Support Vector Machines

Fatimah Othman Hamarasool[1*], Delshad Shaker Ismael Botani[2]

[1]Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University-Erbil, Erbil, Kurdistan Region, Iraq. E-mail: fatimah.hamarasool@su.edu.krd
[2]Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University-Erbil, Erbil, Kurdistan Region, Iraq. E-mail: delshd.botani@su.edu.krd

Abstract
Ordinal classification of global CO_2 emission growth requires models that accommodate ordering, nonlinearity, imbalance, and temporal dependence. This study proposes a probability-level hybrid framework combining ordinal logistic regression (OLR) and a calibrated support vector machine (SVM), using fixed equal-weight fusion (HSM) and adaptive ordinal attention (AOA–HSM). Performance was assessed through 9,000 simulations spanning linear, mixed, and nonlinear mechanisms at n=100,300,500, and through 59 annual observations (1965–2023) with expanding-window temporal validation yielding 30 held-out predictions. OLR provided the strongest ordinal discrimination in linear and mixed settings, whereas HSM was generally more stable; AOA–HSM improved probability quality in selected nonlinear settings. Empirically, calibrated SVM achieved the highest accuracy (0.667), ordinal random forest the highest QWK (0.441), and AOA–HSM the best LogLoss (0.855) and Brier score (0.515). No method dominated all criteria, indicating that adaptive probability fusion is most useful when probabilistic reliability under nonlinear heterogeneity is prioritized.

Keywords: Ordinal classification; Machine learning; Ordinal logistic regression; Calibrated support vector machine; Adaptive probability fusion; Quadratic weighted kappa; CO_2 emission growth; Temporal validation.

Introduction

Annual changes in global carbon dioxide (CO_2) emissions are commonly summarised using ordered categories representing low, moderate, and high emission growth. Such categories can support environmental monitoring, comparative assessment, and policy-oriented interpretation. Because the categories possess a natural order but the distances between them are not necessarily equal, their statistical analysis is appropriately formulated as an ordinal classification problem (United Nations Framework Convention on Climate Change 2015; Intergovernmental Panel on Climate Change 2023; Agresti 2010). Developing models that preserve this ordering while accommodating nonlinear predictor effects, class imbalance, and temporal dependence remains an important challenge in environmental data analysis.

Ordinal classification arises whenever the response consists of ranked categories, including severity levels, risk classes, and ordered policy indicators. Methods developed for nominal classification may fail to distinguish between adjacent and severe classification errors. For example, predicting a high emission-growth level as moderate is less serious than predicting it as low. Consequently, model evaluation should incorporate ordinal-sensitive measures in addition to conventional classification indices.

Ordinal Logistic Regression (OLR) is among the most established approaches for ordered outcomes. Under the cumulative-link formulation, OLR represents the response through ordered thresholds and a common regression structure (McCullagh1980; Agresti 2010). Its principal advantages are interpretability, direct estimation of class probabilities, and explicit use of the response ordering. These features are particularly valuable in environmental applications, where transparent model interpretation may be as important as predictive performance. However, the proportional-odds assumption and linear predictor structure may be restrictive when environmental relationships contain nonlinearities, interactions, or heterogeneous effects (Harrell 2015; Stern 2016; He and Garcia 2009; Pachauri et al. 2014). Support Vector Machines (SVMs) provide a complementary modelling strategy through margin-based learning and kernel transformations (Cortes and Vapnik 1995; Scholkopf and Smola 2002). Their ability to capture nonlinear decision boundaries has motivated their use in complex environmental classification problems. Nevertheless, standard multiclass SVMs treat response categories as nominal and do not directly exploit their ordering. They also require an additional calibration procedure when class-probability estimates are needed. Ordinal SVM extensions have addressed some of these limitations through ranking constraints, threshold formulations, and data-replication strategies (Herbrich et al. 2000; Shashua and Levin 2003; Chu and Keerthi 2005; Cardoso and da Costa 2007). However,

such formulations may be less straightforward to interpret and do not always provide calibrated multiclass probabilities.

Hybrid and ensemble methods offer a possible means of combining the complementary strengths of statistical and machine-learning models (Dietterich 2000; Zhou 2012). Probability-level fusion is particularly attractive because predictions from heterogeneous models can be combined while retaining a valid probability distribution. Recent studies have also proposed boosted, contrastive, and optimization-based approaches for ordinal learning (Sharabiani et al. 2025; Pitawela et al. 2025; Bala´c et al. 2025). However, many existing hybrid methods use fixed global weights, focus mainly on nominal performance measures, or provide limited evidence on how their behaviour changes across linear and nonlinear data structures.

Global CO_2 emissions growth data present several modelling challenges simultaneously. The number of annual observations is limited, the response categories are imbalanced, predictor effects may be nonlinear, and temporal dependence may invalidate randomly shuffled train–test evaluation. These characteristics require a framework that combines interpretable ordinal modelling with nonlinear predictive flexibility and that is evaluated using time-respecting procedures. They also require cautious interpretation because a more flexible hybrid model is not guaranteed to outperform simpler base learners in every sample or under every performance measure.

To address these issues, this study develops a probability-level hybrid framework combining OLR and a probability-calibrated SVM. The fixed Hybrid Model (HSM) averages the two probability vectors using equal weights. Its adaptive extension, AOA–HSM, modifies the SVM contribution according to the observation-specific Euclidean disagreement between the two base-model probability distributions. The adaptive sensitivity parameter is selected using inner-validation quadratic weighted kappa, with ranked probability score used for tie-breaking. The resulting fusion remains a valid probability distribution, although probability fusion alone does not guarantee complete ordinal coherence or predictive superiority.

The framework is examined through three complementary forms of evidence. First, its probability validity, limiting behaviour, attention bounds, and continuity are established analytically. Second, a simulation study evaluates OLR, Calibrated-SVM, HSM, and AOA–HSM under linear, nonlinear, and mixed data-generating mechanisms, three sample sizes, and 1000 replications per configuration. Third, the empirical application uses annual global CO_2 emission-growth data from 1965–2023 under a leakage-free expanding-window design. Ordinal gradient boosting and ordinal random forest are included as additional empirical comparators, and uncertainty is assessed using a moving-block bootstrap.

The study does not assume that the adaptive model must dominate all alternatives. Instead, it evaluates when fixed or adaptive fusion is beneficial and whether gains occur in ordinal discrimination, classification accuracy, or probability quality. This distinction is important because the simulations and empirical application indicate that model rankings may differ across evaluation criteria and underlying data structures.

Methodology

This section formulates the four models used throughout the simulation study and empirical application: ordinal logistic regression (OLR), probability-calibrated support vector machine (Calibrated-SVM), a fixed hybrid probability-fusion model (HSM), and its adaptive extension (AOA–HSM). The same fitted and calibrated SVM is used both as a standalone comparator and as the SVM component of the two hybrid models.

2.1 Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of explanatory variables and $y_i \in \{1, \dots, K\}$, $1 \prec 2 \prec \cdots \prec K$, is an ordinal response. For a predictor vector $\mathbf{x}$, define the class-probability vector

$$\boldsymbol{\pi}(\mathbf{x}) = \big(\pi_1(\mathbf{x}), \dots, \pi_K(\mathbf{x})\big)^{\top},$$

where

$$\pi_k(\mathbf{x}) = P(Y = k \mid \mathbf{X} = \mathbf{x}).$$

A valid probability vector belongs to the simplex

$$\Delta^{K-1} = \left\{\boldsymbol{\pi} \in [0,1]^K : \sum_{k=1}^{K} \pi_k = 1\right\}.$$

For all models, the predicted response category is

$$\hat{Y}(\mathbf{x}) = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}}\ \hat{\pi}_k(\mathbf{x}).$$

Independent sampling is assumed for simulation datasets generated independently. The empirical observations form a temporally ordered sequence and are therefore evaluated using chronological training and test periods.

2.2 Ordinal Logistic Regression

OLR was used as the interpretable ordinal base learner. Cumulative-link models provide a principled framework for modelling ordered categorical responses without treating the category labels as equally spaced numerical values (McCullagh 1980). Under the cumulative-logit proportional-odds model,

$$\log\left[\frac{P(Y \le k \mid \mathbf{x})}{P(Y > k \mid \mathbf{x})}\right] = \theta_k - \mathbf{x}^\top \boldsymbol{\beta}, \qquad k = 1, \dots, K-1,$$

where

$$\theta_1 < \theta_2 < \cdots < \theta_{K-1}$$

are ordered cut-points and $\boldsymbol{\beta}$ is the common coefficient vector.
Let

$$\hat{F}_{k,\mathrm{OLR}}(\mathbf{x}) = \hat{P}_{\mathrm{OLR}}(Y \le k \mid \mathbf{x})$$

denote the estimated cumulative probabilities. The corresponding class probabilities are

$$\hat{\pi}_{1,\mathrm{OLR}}(\mathbf{x}) = \hat{F}_{1,\mathrm{OLR}}(\mathbf{x}).$$
$$\hat{\pi}_{k,\mathrm{OLR}}(\mathbf{x}) = \hat{F}_{k,\mathrm{OLR}}(\mathbf{x}) - \hat{F}_{k-1,\mathrm{OLR}}(\mathbf{x}), \quad k = 2, \dots, K-1.$$
$$\hat{\pi}_{K,\mathrm{OLR}}(\mathbf{x}) = 1 - \hat{F}_{K-1,\mathrm{OLR}}(\mathbf{x}).$$

Hence,

$$\hat{\boldsymbol{\pi}}_{\mathrm{OLR}}(\mathbf{x}) \in \Delta^{K-1}.$$

OLR explicitly represents the ordering of the response and provides interpretable coefficient effects. Its principal restriction is the common-slope proportional-odds structure, which may be inadequate when predictor effects differ across cumulative response thresholds (McCullagh 1980; Brant 1990).
The proportional-odds assumption was assessed in the empirical application using a global test of parallel lines. This test evaluates whether the slope coefficients remain constant across the cumulative logits (Brant, 1990). It was applied to the full-sample cumulative-logit model using the same predictor coding as in the reported OLR analysis.

2.3Probability-Calibrated Support Vector Machine

A support vector machine (SVM) was used as the flexible base learner. For each binary classification subproblem, let z_i∈{-1,+1} denote the binary class label. The class-weighted soft-margin SVM constructs a maximum-margin decision boundary while penalising observations that violate the margin constraints (Cortes and Vapnik 1995; Burges 1998). It solves

$$\min_{\mathbf{v},b,\boldsymbol{\xi}} \left\{ \frac{1}{2} \| \mathbf{v} \|^2 + C \sum_{i=1}^{N} \omega_i \xi_i \right\},$$

subject to

$$z_i[\mathbf{v}^\top \phi(\mathbf{x}_i) + b] \ge 1 - \xi_i, \qquad \xi_i \ge 0,$$

where $C > 0$ controls the overall penalty for margin violations, ω_i is the class-specific balancing weight, and $\phi(\cdot)$ denotes the kernel-induced feature mapping.
For nonlinear classification, the radial basis function kernel was defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right],$$

where $\gamma > 0$ controls the locality of the decision boundary. The RBF kernel permits nonlinear separation by expressing similarity as a decreasing function of the distance between predictor vectors (CortesVapnik1995; Burges1998).
For the K-class response, the nominal multiclass SVM was implemented using a one-versus-one decomposition, in which one binary classifier is constructed for every pair of response classes (Hsu and Lin 2002; Chang and Lin 2011). Consequently,

$$\frac{K(K-1)}{2}$$

binary classifiers were fitted. For the present three-class problem, this corresponded to three pairwise classifiers. Class-balanced weights were used to reduce the influence of response imbalance.
Raw SVM decision scores are not calibrated class probabilities. Therefore, the SVM was wrapped in a cross-validated sigmoid calibration procedure based on Platt's logistic transformation (Platt 1999; Lin et al. 2007). Calibration used at most three stratified folds, with the number of folds reduced when required by the smallest training-class count.
The resulting calibrated probability vector was

$$\hat{\boldsymbol{\pi}}_{\mathrm{SVM}}(\mathbf{x}) = \left(\hat{\pi}_{1,\mathrm{SVM}}(\mathbf{x}), \dots, \hat{\pi}_{K,\mathrm{SVM}}(\mathbf{x})\right)^\top \in \Delta^{K-1}.$$

Predictor standardisation, kernel and hyperparameter selection, SVM fitting, and probability calibration were conducted entirely within the relevant training data, without access to the corresponding outer-test observations.
The calibrated probabilities were aligned to the common class order $\{1, \dots, K\}$, and the predicted category was obtained as

$$\hat{Y}(\mathbf{x}) = \arg\max_{k \in \{1,\dots,K\}} \hat{\pi}_{k,\mathrm{SVM}}(\mathbf{x}).$$

The same calibrated probability vector $\hat{\boldsymbol{\pi}}_{\mathrm{SVM}}(\mathbf{x})$ was used unchanged for the standalone Calibrated-SVM, HSM, and AOA–HSM models.

2.4 Fixed Hybrid Probability-Fusion Model

The fixed hybrid model, denoted by HSM, combines the two base-model probability vectors using equal weights. Linear combination and sum rules are established approaches for combining the outputs of multiple classifiers (Kittler et al. 1998). In the present study, equal weighting was adopted as a transparent and parsimonious fusion baseline:

$$\hat{\boldsymbol{\pi}}_{\mathrm{HSM}}(\mathbf{x}) = \frac{1}{2}\hat{\boldsymbol{\pi}}_{\mathrm{OLR}}(\mathbf{x}) + \frac{1}{2}\hat{\boldsymbol{\pi}}_{\mathrm{SVM}}(\mathbf{x}).$$

The equal-weight rule does not require estimation of an additional mixture parameter. Technically, HSM is a probability-averaging model rather than classical stacked generalisation, because stacking requires a second-level learner trained to combine the outputs of the base learners (Wolpert 1992).

2.5 Adaptive Ordinal Attention Hybrid Model

AOA–HSM is the adaptive probability-fusion rule proposed in this study. It extends HSM by allowing the contribution of the calibrated SVM to vary across observations according to the disagreement between the two base-model probability vectors.

$$\delta(\mathbf{x}) = \|\hat{\boldsymbol{\pi}}_{\mathrm{SVM}}(\mathbf{x}) - \hat{\boldsymbol{\pi}}_{\mathrm{OLR}}(\mathbf{x})\|_2.$$

Because both vectors belong to the probability simplex,

$$0 \le \delta(\mathbf{x}) \le \sqrt{2}.$$

The observation-specific SVM weight is

$$\alpha(\mathbf{x};\tau) = \frac{1}{1+\exp[-2\tau\delta(\mathbf{x})]}, \qquad \tau \ge 0,$$

where τ controls the sensitivity of the fusion rule to base-model disagreement.

The adaptive probability vector is

$$\hat{\boldsymbol{\pi}}_{\mathrm{AOA}}(\mathbf{x};\tau) = \alpha(\mathbf{x};\tau)\hat{\boldsymbol{\pi}}_{\mathrm{SVM}}(\mathbf{x}) + [1-\alpha(\mathbf{x};\tau)]\hat{\boldsymbol{\pi}}_{\mathrm{OLR}}(\mathbf{x}).$$

When τ=0, the two models receive equal weight. For τ>0, larger disagreement increases the contribution of the SVM. The rule is therefore directional: it moves from equal weighting toward the nonlinear SVM component. It does not estimate which base model is locally more accurate, and predictive improvement is not guaranteed. The resulting mechanism should therefore be interpreted as a directional adaptive fusion rule rather than an estimator of local base-model reliability. Predictive improvement over fixed probability averaging is not theoretically guaranteed and was consequently evaluated empirically.

2.6 Selection of the Adaptive Parameter

Let $\mathcal{T} = \{\tau_1, \ldots, \tau_L\}$ denote the candidate grid. For each candidate value, AOA–HSM probabilities were generated from out-of-fold OLR and Calibrated-SVM probabilities within the training data. Parameter selection followed the ordered rule:

1. maximise inner-validation quadratic weighted kappa;
2. among tied candidates, minimise the ranked probability score;
3. if a tie remained, select the smallest value of τ.

Equivalently,

$$\hat{\tau} = \arg\operatorname*{lex\,opt}_{\tau\in\mathcal{T}}\{-\mathrm{QWK}(\tau), \mathrm{RPS}(\tau), \tau\}.$$

The candidate set included $\tau = 0$, allowing the adaptive procedure to recover HSM when the training data did not support additional reweighting.

2.7 Analytical Properties

Proposition 1 (Validity and limiting behaviour).

Suppose

$$\hat{\boldsymbol{\pi}}_{\mathrm{OLR}}(\mathbf{x}), \hat{\boldsymbol{\pi}}_{\mathrm{SVM}}(\mathbf{x}) \in \Delta^{K-1}.$$

Then, for every finite $\tau \ge 0$*:*

1. $\hat{\boldsymbol{\pi}}_{\mathrm{HSM}}(\mathbf{x})$ *and* $\hat{\boldsymbol{\pi}}_{\mathrm{AOA}}(\mathbf{x};\tau)$ *belong to* Δ^{K-1}*;*
2. $\frac{1}{2} \le \alpha(\mathbf{x};\tau) \le \frac{1}{1+\exp(-2\sqrt{2}\tau)} < 1$;
3. $\hat{\boldsymbol{\pi}}_{\mathrm{AOA}}(\mathbf{x};0) = \hat{\boldsymbol{\pi}}_{\mathrm{HSM}}(\mathbf{x})$;
4. *the adaptive weight is non-decreasing in both* τ *and* δ*.*

Proof.

Both fusion rules are convex combinations of two valid probability vectors and therefore have non-negative components summing to one.

Since

$$0 \le \delta(\mathbf{x}) \le \sqrt{2},$$

application of the monotone logistic function gives the stated bounds for α. When $\tau = 0$,

$$\alpha(\mathbf{x}; 0) = \frac{1}{2},$$

which reduces AOA–HSM exactly to HSM. Finally,

$$\frac{\partial \alpha}{\partial \tau} = 2\delta\alpha(1-\alpha) \geq 0, \qquad \frac{\partial \alpha}{\partial \delta} = 2\tau\alpha(1-\alpha) \geq 0.$$

Proposition 2 (Continuity).
If $\hat{\boldsymbol{\pi}}_{\mathrm{OLR}}(\mathbf{x})$ and $\hat{\boldsymbol{\pi}}_{\mathrm{SVM}}(\mathbf{x})$ are continuous in $\mathbf{x}$, then $\hat{\boldsymbol{\pi}}_{\mathrm{AOA}}(\mathbf{x}; \tau)$ is continuous for every finite τ.
Proof.
The Euclidean distance between the two probability vectors is continuous, the logistic transformation is continuous, and the final fusion consists of sums and products of continuous functions.
These results establish probability validity and smoothness of the fusion rule. They do not imply lower prediction error or superiority over either base learner.

2.8 Ordinal Interpretation

The proposed framework is ordinally oriented through the cumulative-link formulation of OLR (McCullagh 1980), the use of quadratic weighted kappa (Cohen 1968), and evaluation with ordinal error and probability measures, including MAE, MSE, and RPS (Epstein 1969; Gneiting and Raftery 2007). More specifically, ordinality enters the framework through:

1. the cumulative-link structure of the OLR component;
2. the ordered response categories;
3. selection of τ using QWK;
4. evaluation using ordinal-sensitive measures, including QWK, MAE, MSE, and RPS.

Experimental Design

The proposed framework was evaluated using a controlled simulation study and an empirical application to annual global CO_2 emission-growth data. In both experiments, preprocessing, probability calibration, hyperparameter selection, and model fitting were conducted using training data only.

3.1 Simulation Study

3.1.1 Data-Generating Mechanisms

Three data-generating mechanisms were considered: linear, nonlinear, and mixed. For each mechanism,

$$n \in \{100,300,500\}$$

was examined, with 1000 independent replications per sample size. The complete design therefore comprised

$$3 \times 3 \times 1000 = 9000$$

retained datasets.
Each dataset contained six correlated continuous predictors,

$$\mathbf{X}_i \sim N_6(\mathbf{0}, \boldsymbol{\Sigma}), \qquad \Sigma_{j\ell} = 0.30^{|j-\ell|}.$$

The linear component was

$$g_{\mathrm{L}}(\mathbf{X}) = 1.20X_1 - 1.00X_2 + 0.80X_3 - 0.60X_4 + 0.40X_5 - 0.30X_6,$$

and the nonlinear component was

$$g_{\mathrm{N}}(\mathbf{X}) = \sin(X_1) + 0.70(X_2^2 - 1) - 0.80X_3X_4 + 0.60\tanh(X_5) - 0.50\left(|X_6| - \sqrt{\frac{2}{\pi}}\right).$$

The scenario-specific signal was defined as

$$g_s(\mathbf{X}) = \begin{cases} g_{\mathrm{L}}(\mathbf{X}), & s = \text{linear}, \\ g_{\mathrm{N}}(\mathbf{X}), & s = \text{nonlinear}, \\ g_{\mathrm{L}}(\mathbf{X}) + g_{\mathrm{N}}(\mathbf{X}), & s = \text{mixed}. \end{cases}$$

To ensure comparable signal strength, each scenario was scaled to have standard deviation 1.50 using an independent calibration population of 100,000 observations. The latent response was

$$Z_i = a_s g_s(\mathbf{X}_i) + \varepsilon_i, \qquad \varepsilon_i \sim \text{Logistic}(0,1),$$

where

$$a_s = \frac{1.50}{\mathrm{SD}[g_s(\mathbf{X})]}.$$

Scenario-specific cut-points were fixed at the 20th and 80th percentiles of the calibration-population latent response:

$$Y_i = \begin{cases} 1, & Z_i \leq c_{1s}, \\ 2, & c_{1s} < Z_i \leq c_{2s}, \\ 3, & Z_i > c_{2s}. \end{cases}$$

This produced approximate class proportions of 0.20, 0.60, and 0.20. A sample was regenerated when one or more response levels were absent or when any level contained fewer than six observations, with a maximum of 100 attempts.

3.1.2 Models, Validation, and Hyperparameter Selection

Four models were compared in the simulation study: OLR, Calibrated-SVM, HSM, and AOA–HSM. Identical generated datasets and cross-validation partitions were used for all models within each replication. Performance was estimated using nested stratified cross-validation with three outer folds and two inner folds. The outer-test predictions were pooled to form one complete out-of-fold prediction vector per replication, and individual folds were not treated as independent observations. All preprocessing, model fitting, calibration, and hyperparameter selection were restricted to the corresponding training data.

The multiclass SVM used a one-versus-one SVC with class-balanced weights and linear or RBF kernels. The tuning grid was

$$C \in \{0.01,0.1,1,10,100\},$$

and, for the RBF kernel,

$$\gamma \in \{\text{scale}, 0.001,0.01,0.1,1\}.$$

Class probabilities were obtained using cross-validated sigmoid calibration with up to three stratified folds, reduced when required by the smallest training-class count. For AOA–HSM,

$$\tau \in \{0,0.05,0.10,0.25,0.50,1,2,3\},$$

where $\tau = 0$ allowed exact recovery of HSM. Hyperparameters were selected by maximising inner-validation QWK, with the lowest RPS used to resolve ties and the smallest τ selected when a tie remained.

3.1.3 Performance Measures and Statistical Comparison

Predictive performance was evaluated using six complementary measures covering conventional classification accuracy, class-balanced performance, ordinal agreement, ordinal error, and probabilistic accuracy. These measures were accuracy, balanced accuracy (BACC), Macro-F1 score, quadratic weighted kappa (QWK), ordinal mean absolute error (MAE), and ranked probability score (RPS) (Sokolova and Lapalme 2009; Brodersen et al. 2010). For each model and simulation configuration, the mean and standard deviation of each performance measure were calculated across the 1,000 independent replications. Higher values indicate better performance for accuracy, BACC, Macro-F1, and QWK, whereas lower values indicate better performance for MAE and RPS.

Statistical comparisons were conducted separately for each combination of simulation scenario, sample size, and performance measure. Because all four competing models were evaluated on the same simulated dataset within each replication, each replication was treated as a matched block. The Friedman test was first used to assess overall differences among the models (Friedman 1937). When the Friedman test indicated a statistically significant overall difference, the six possible pairwise comparisons were subsequently performed using the Wilcoxon signed-rank test (Wilcoxon 1945). Holm's sequential procedure was then applied to the six pairwise p-values separately within each scenario, sample-size, and performance-measure combination to control the family-wise error rate (Holm 1979). Statistical significance was assessed at the$\alpha = 0.05$ level.

3.2 Empirical Application

3.2.1 Data and Ordinal Response

The empirical dataset contained 59 annual observations covering 1965–2023 the 1964 CO_2 emission value was used only as the baseline required to calculate the growth rate for 1965. The predictors were global population, annual GDP growth, annual rainfall, renewable-energy share, and a three-level fossil-fuel category. Population and GDP data were obtained from the World Development Indicators, whereas renewable-energy and CO_2 emission data were obtained from Our World in Data and the Global Carbon Budget. Detailed definitions, measurement units, and source information are provided in Table 1.

Table 1. Empirical variables, units, roles, and principal sources.

Variable	Definition	Unit	Role	Source
Population	Annual global population	persons	continuous predictor	World Development Indicators
GDP growth	Annual global GDP growth	%	continuous predictor	World Development Indicators
Rainfall	Annual global rainfall series used in the archived dataset	mm/year	continuous predictor	archived data dictionary

Variable	Definition	Unit	Role	Source
Renewable-energy share	Global renewable-energy share	%	continuous predictor	Our World in Data
Fossil-fuel category	Prespecified three-level fossil-fuel indicator; Level 1 was the reference category	category	categorical predictor	derived variable; archived construction code
CO_2 emissions	Annual global emissions used only to construct the response	tonnes/year	response construction	Our World in Data/Global Carbon Budget
Emission-growth level	Level 1: $G_t < 1\%$; Level 2: $1\% \leq G_t < 5\%$; Level 3: $G_t \geq 5\%$	ordinal level	response	derived

Let E_t denote global CO_2 emissions in year t. Annual emission growth was calculated as

$$G_t = 100\left(\frac{E_t - E_{t-1}}{E_{t-1}}\right).$$

The ordinal response was defined by

$$Y_t = \begin{cases} 1, & G_t < 1\%, \\ 2, & 1\% \leq G_t < 5\%, \\ 3, & G_t \geq 5\%. \end{cases}$$

The resulting class counts were 17, 34, and 8, respectively. These thresholds were treated as prespecified operational cut-offs rather than internationally mandated standards.
Calendar year, annual CO_2 emissions, and the continuous growth rate were excluded from the predictor set to prevent temporal indexing and outcome leakage.

3.2.2 Temporal Validation

Temporal dependence was assessed using the sample autocorrelation function and Ljung–Box tests for the continuous growth series, together with lag-one Spearman correlation and a transition-based chi-square test for the ordinal response. Because temporal dependence could not be excluded, the primary empirical evaluation used three expanding-window splits that preserved chronological ordering.

Table 2. Expanding-window validation design and class distributions.

Test period	Training years	N_{train}	Training classes			N_{test}	Test classes		
			L1	L2	L3		L1	L2	L3
1994–2003	1965–1993	29	7	17	5	10	3	6	1
2004–2013	1965–2003	39	10	23	6	10	2	7	1
2014–2023	1965–2013	49	12	30	7	10	5	4	1
Full sample	1965–2023	59	17	34	8	–	–	–	–
Pooled test	–	–	–	–	–	30	10	17	3

The three outer-test periods yielded 30 chronologically held-out predictions covering 1994–2023. Performance measures were calculated from these pooled out-of-sample predictions rather than from fitted or randomly shuffled observations.

3.2.3 Preprocessing and Inner Selection

Within every historical outer-training window, continuous predictors were standardised using training-set means and standard deviations. The fossil-fuel category was one-hot encoded using Level 1 as the reference category, and effectively constant columns were removed before fitting OLR. No information from the corresponding outer-test period was used during preprocessing, model fitting, calibration, or hyperparameter selection.

Because the historical training windows contained few Level-3 observations, the primary analysis used stratified inner cross-validation with at most three folds. The number of folds was restricted by the smallest training-class count. This procedure maintained complete separation from the outer-test period, although it did not preserve the temporal order within the historical training window.

To evaluate the sensitivity of the results to this choice, all hyperparameter selection was repeated using fully chronological inner validation. The chronological analysis was treated as a conservative sensitivity assessment because some early validation windows contained very few, or no, Level-3 observations. Results from both tuning designs are reported separately.

3.2.4 Comparators and Hyperparameter Search

Six models were evaluated in the empirical analysis: OLR, Calibrated-SVM, HSM, AOA–HSM, ordinal gradient boosting (OGB), and ordinal random forest (ORF). OLR was fitted as a cumulative-logit proportional-odds model and required no hyperparameter tuning.

For the Calibrated-SVM, both linear and radial basis function (RBF) kernels were considered. The regularization parameter was selected from $C \in \{0.01,0.1,1,10,100\}$ and, for the RBF kernel, the kernel coefficient was selected from $\gamma \in \{\text{scale},0.001,0.01,0.1,1\}$. Class-balanced weights were used, and cross-validated sigmoid calibration was applied to obtain class-probability estimates. HSM combined the OLR and Calibrated-SVM probability vectors with equal weights and therefore required no additional tuning. For AOA–HSM, the adaptive parameter was selected from $\tau \in \{0,0.05,0.10,0.25,0.50,1,2,3\}$.

OGB and ORF were implemented through cumulative binary models for $P(Y>1)$ and $P(Y>2)$. To preserve the ordinal structure, estimated cumulative probabilities were adjusted, when necessary, to satisfy $P(Y>1) \geq P(Y>2)$ before the corresponding class probabilities were recovered. For OGB, the candidate grid included 100 or 200 trees, learning rates of 0.03 or 0.10, tree depths of 1 or 2, minimum leaf sizes of 2 or 4, and subsampling fractions of 0.8 or 1.0. For ORF, the number of trees was selected from {300,600}, maximum depth from {3,5,None}, minimum leaf size from {1,2,4}, and the maximum-feature option from {sqrt,1.0}.

For all tuned models, hyperparameters were selected within the training data by maximising inner-validation QWK. Ties were resolved using the lowest RPS; for AOA–HSM, any remaining tie was resolved by selecting the smallest value of τ. The selected settings varied across the three temporal evaluation periods. For 1994–2003, the Calibrated-SVM used an RBF kernel with $C=10$ and $\gamma=0.1$, while AOA–HSM selected $\tau=0.50$. For 2004–2013, a linear SVM with $C=1$ was selected and $\tau=3.00$. For 2014–2023, the selected SVM again used an RBF kernel with $C=10$ and $\gamma=0.1$, with $\tau=0.05$.

The selected OGB configurations used 100 trees with a learning rate of 0.03 and depth 1 in all three periods. The minimum leaf size and subsampling fraction were (2,0.8) for 1994–2003, (4,1.0) for 2004–2013, and (2,1.0) for 2014–2023. For ORF, the selected configurations used 600 trees, depth 3, minimum leaf size 4, and maximum features 1.0 in the first two periods; for 2014–2023, the corresponding settings were 300 trees, depth 3, minimum leaf size 2, and maximum features 1.0.

3.2.5- Performance, Uncertainty, and Sensitivity

Empirical performance was assessed using accuracy, balanced accuracy, Macro-F1, QWK, MAE, MSE, Spearman correlation, multiclass AUC, LogLoss, and the multiclass Brier score.

Uncertainty was quantified using a moving-block bootstrap with

$B=2000$

replications and a block length of five consecutive years. Identical bootstrap blocks were used across models to preserve paired comparisons. Percentile-based 95% confidence intervals were reported; AUC was calculated only when all three response levels were present in the bootstrap sample.

Two substantive sensitivity analyses were performed. First, the response thresholds were changed from 1% and 5% to 0.5% and 4.5%. Second, hyperparameter selection was repeated using fully chronological inner validation. Computational time was also recorded for each model and reported separately from predictive performance.

Results

4.1 Simulation Results

The four models were evaluated under linear, mixed, and nonlinear data-generating mechanisms at sample

sizes n=100, 300, and 500, with 1,000 independent replications for each configuration. Table 3 summarises performance in terms of accuracy, balanced accuracy (BACC), Macro-F1, quadratic weighted kappa (QWK), ordinal mean absolute error (MAE), and ranked probability score (RPS). Results are reported as replication-level means and standard deviations. Higher values indicate better performance for accuracy, BACC, Macro-F1, and QWK, whereas lower values are preferable for MAE and RPS.

Table 3. Multimetric simulation performance across the nine scenario–sample-size configurations. Results are mean (standard deviation) across 1,000 independent replications. Larger values are preferable for Accuracy, BACC, Macro-F1, and QWK; smaller values are preferable for MAE and RPS.

Scenario	*n*	Model	Accuracy	BACC	Macro-F1	QWK	MAE	RPS
Linear	100	OLR	0.643 (0.053)	0.511 (0.069)	0.529 (0.077)	0.413 (0.115)	0.3629 (0.0563)	0.1303 (0.0156)
Linear	100	Calibrated-SVM	0.600 (0.050)	0.352 (0.036)	0.292 (0.057)	0.065 (0.099)	0.4027 (0.0522)	0.1459 (0.0134)
Linear	100	HSM	0.644 (0.050)	0.462 (0.069)	0.470 (0.089)	0.327 (0.135)	0.3590 (0.0516)	0.1303 (0.0138)
Linear	100	AOA–HSM	0.641 (0.050)	0.453 (0.071)	0.456 (0.093)	0.309 (0.142)	0.3614 (0.0515)	0.1311 (0.0139)
Linear	300	OLR	0.659 (0.028)	0.520 (0.041)	0.546 (0.045)	0.432 (0.067)	0.3444 (0.0288)	0.1234 (0.0077)
Linear	300	Calibrated-SVM	0.605 (0.026)	0.357 (0.038)	0.299 (0.062)	0.076 (0.101)	0.3962 (0.0271)	0.1385 (0.0078)
Linear	300	HSM	0.652 (0.026)	0.468 (0.051)	0.483 (0.065)	0.341 (0.099)	0.3497 (0.0268)	0.1264 (0.0074)
Linear	300	AOA–HSM	0.652 (0.026)	0.466 (0.051)	0.480 (0.066)	0.337 (0.100)	0.3503 (0.0268)	0.1266 (0.0074)
Linear	500	OLR	0.662 (0.021)	0.523 (0.030)	0.551 (0.033)	0.439 (0.049)	0.3409 (0.0218)	0.1221 (0.0059)
Linear	500	Calibrated-SVM	0.613 (0.021)	0.372 (0.038)	0.327 (0.061)	0.116 (0.098)	0.3884 (0.0211)	0.1343 (0.0058)
Linear	500	HSM	0.656 (0.021)	0.479 (0.041)	0.500 (0.050)	0.364 (0.077)	0.3460 (0.0214)	0.1248 (0.0056)
Linear	500	AOA–HSM	0.656 (0.021)	0.478 (0.041)	0.498 (0.051)	0.362 (0.077)	0.3463 (0.0214)	0.1249 (0.0056)
Mixed	100	OLR	0.607 (0.054)	0.445 (0.065)	0.452 (0.080)	0.292 (0.127)	0.4020 (0.0572)	0.1441 (0.0163)
Mixed	100	Calibrated-SVM	0.596 (0.050)	0.344 (0.026)	0.279 (0.045)	0.042 (0.076)	0.4064 (0.0526)	0.1505 (0.0140)
Mixed	100	HSM	0.613 (0.050)	0.403 (0.057)	0.388 (0.083)	0.204 (0.133)	0.3910 (0.0527)	0.1410 (0.0145)
Mixed	100	AOA–HSM	0.612 (0.050)	0.394 (0.057)	0.371 (0.084)	0.181 (0.135)	0.3921 (0.0526)	0.1417 (0.0145)
Mixed	300	OLR	0.620 (0.030)	0.442 (0.038)	0.452 (0.049)	0.294 (0.075)	0.3849 (0.0309)	0.1376 (0.0082)

Scenario	n	Model	Accuracy	BACC	Macro-F1	QWK	MAE	RPS
Mixed	300	Calibrated-SVM	0.600 (0.028)	0.343 (0.019)	0.274 (0.035)	0.036 (0.058)	0.4003 (0.0285)	0.1450 (0.0077)
Mixed	300	HSM	0.618 (0.027)	0.395 (0.038)	0.375 (0.060)	0.184 (0.093)	0.3839 (0.0281)	0.1375 (0.0077)
Mixed	300	AOA–HSM	0.617 (0.027)	0.392 (0.038)	0.370 (0.062)	0.177 (0.096)	0.3845 (0.0279)	0.1377 (0.0078)
Mixed	500	OLR	0.623 (0.022)	0.446 (0.030)	0.458 (0.039)	0.304 (0.060)	0.3815 (0.0234)	0.1363 (0.0062)
Mixed	500	Calibrated-SVM	0.601 (0.021)	0.352 (0.022)	0.291 (0.041)	0.062 (0.066)	0.3995 (0.0219)	0.1419 (0.0062)
Mixed	500	HSM	0.620 (0.021)	0.406 (0.033)	0.393 (0.051)	0.211 (0.078)	0.3814 (0.0218)	0.1361 (0.0059)
Mixed	500	AOA–HSM	0.620 (0.021)	0.404 (0.033)	0.390 (0.052)	0.207 (0.080)	0.3818 (0.0217)	0.1362 (0.0059)
Nonlinear	100	OLR	0.579 (0.056)	0.363 (0.042)	0.335 (0.063)	0.086 (0.112)	0.4359 (0.0632)	0.1623 (0.0174)
Nonlinear	100	Calibrated-SVM	0.597 (0.051)	0.339 (0.018)	0.268 (0.032)	0.022 (0.053)	0.4060 (0.0537)	0.1558 (0.0144)
Nonlinear	100	HSM	0.594 (0.053)	0.348 (0.028)	0.292 (0.047)	0.047 (0.080)	0.4115 (0.0574)	0.1550 (0.0152)
Nonlinear	100	AOA–HSM	0.596 (0.053)	0.345 (0.025)	0.285 (0.044)	0.040 (0.072)	0.4090 (0.0567)	0.1548 (0.0150)
Nonlinear	300	OLR	0.593 (0.032)	0.348 (0.018)	0.296 (0.035)	0.051 (0.056)	0.4120 (0.0339)	0.1553 (0.0090)
Nonlinear	300	Calibrated-SVM	0.601 (0.029)	0.339 (0.012)	0.264 (0.024)	0.020 (0.039)	0.4001 (0.0298)	0.1513 (0.0081)
Nonlinear	300	HSM	0.600 (0.030)	0.338 (0.011)	0.265 (0.022)	0.019 (0.034)	0.4014 (0.0308)	0.1513 (0.0083)
Nonlinear	300	AOA–HSM	0.600 (0.030)	0.338 (0.010)	0.264 (0.021)	0.018 (0.033)	0.4007 (0.0304)	0.1510 (0.0083)
Nonlinear	500	OLR	0.595 (0.024)	0.344 (0.013)	0.285 (0.026)	0.040 (0.039)	0.4090 (0.0252)	0.1542 (0.0068)
Nonlinear	500	Calibrated-SVM	0.601 (0.022)	0.341 (0.014)	0.269 (0.027)	0.028 (0.043)	0.3996 (0.0226)	0.1478 (0.0065)
Nonlinear	500	HSM	0.600 (0.023)	0.338 (0.009)	0.263 (0.018)	0.017 (0.027)	0.4012 (0.0231)	0.1494 (0.0063)
Nonlinear	500	AOA–HSM	0.600 (0.023)	0.338 (0.010)	0.264 (0.020)	0.019 (0.031)	0.4005 (0.0230)	0.1487 (0.0064)

The results show that no single model dominated across all data-generating mechanisms and evaluation criteria. Under the linear mechanism, OLR consistently provided the strongest ordinal discrimination, achieving the highest QWK at each sample size. Its QWK increased from 0.413 at n=100 to 0.439 at n=500. HSM

achieved similar overall accuracy but consistently higher QWK than AOA–HSM, indicating that additional movement toward the SVM component offered no advantage when the underlying structure was predominantly linear.

A similar pattern was observed under the mixed mechanism. OLR achieved the highest QWK at all three sample sizes, while HSM produced the lowest RPS in each configuration. Both hybrid models nevertheless performed better than the standalone Calibrated-SVM on most ordinal measures, suggesting that incorporating the OLR probability estimates helped stabilise the nominal SVM classifier.

The nonlinear mechanism produced a different pattern. QWK values were low for all four models, reflecting the greater difficulty of the scenario. AOA–HSM showed its clearest advantage in probability quality, achieving the lowest RPS at n=100 and n=300. At n=500, however, Calibrated-SVM achieved the lowest RPS. Differences among Calibrated-SVM, HSM, and AOA–HSM were generally small in this setting.

Across the three mechanisms, replication-level standard deviations generally decreased as sample size increased, indicating greater stability of the performance estimates at larger n. Overall, these results support a scenario-dependent interpretation of model performance rather than uniform superiority of any single approach.

The Friedman test rejected the null hypothesis of equal model ranks for every examined simulation configuration, with

$$61.005 \leq \chi_F^2 \leq 2886.960, \qquad p \leq 3.59 \times 10^{-13}.$$

Because the comparisons were based on 1,000 matched replications, statistical significance was considered together with the magnitude and consistency of the observed performance differences. Table 4 reports the configuration-specific Friedman results for QWK and RPS, together with the Holm-adjusted Wilcoxon signed-rank comparison between HSM and AOA–HSM.

Table 4. Friedman tests and Holm-adjusted HSM–AOA–HSM comparisons for QWK and RPS across the simulation configurations.

Scenario	n	χ_F^2 QWK	p QWK	χ_F^2 RPS	p RPS	p_{Holm} QWK	p_{Holm} RPS
Linear	100	2487.857	$< 10^{-300}$	1825.862	$< 10^{-300}$	7.61×10^{-63}	1.53×10^{-110}
Linear	300	2833.825	$< 10^{-300}$	2767.325	$< 10^{-300}$	5.32×10^{-42}	8.89×10^{-110}
Linear	500	2844.283	$< 10^{-300}$	2873.797	$< 10^{-300}$	6.98×10^{-38}	3.25×10^{-108}
Mixed	100	2066.465	$< 10^{-300}$	1497.402	$< 10^{-300}$	5.78×10^{-65}	8.26×10^{-67}
Mixed	300	2742.836	$< 10^{-300}$	1826.772	$< 10^{-300}$	1.33×10^{-55}	1.48×10^{-98}
Mixed	500	2815.141	$< 10^{-300}$	1733.572	$< 10^{-300}$	7.59×10^{-52}	1.89×10^{-67}
Nonlinear	100	300.135	9.30×10^{-65}	1160.355	2.93×10^{-251}	5.74×10^{-6}	3.21×10^{-5}
Nonlinear	300	495.729	4.02×10^{-107}	1309.553	1.24×10^{-283}	0.547	8.93×10^{-29}
Nonlinear	500	411.371	7.62×10^{-89}	2054.403	$< 10^{-300}$	7.34×10^{-4}	7.26×10^{-113}

Note: Friedman tests were based on 1,000 matched replication-level observations within each scenario and sample-size configuration. The last two columns show the HSM–AOA–HSM Wilcoxon signed-rank comparison after Holm adjustment within the six pairwise model comparisons. Statistical significance was assessed at the α=0.05 level.

The HSM–AOA–HSM comparisons were statistically significant for QWK and RPS in most configurations. The main exception was QWK under the nonlinear mechanism at n=300, where the adjusted comparison was not statistically significant,

p_Holm=0.547.

At n=500, the corresponding QWK difference became statistically significant, but its numerical magnitude remained very small: QWK was 0.019 for AOA–HSM and 0.017 for HSM, with

p_Holm=7.34×10^(-4).

This illustrates why the statistical tests were interpreted together with the magnitude of the observed performance differences rather than from p-values alone.

Fig 1 extends the mean-performance comparison by examining replication-level behaviour. OLR had the highest probability of achieving the best QWK across all nine configurations, although its advantage was less pronounced under the nonlinear mechanism. For RPS, the preferred model depended more strongly on the underlying data structure. OLR and HSM were competitive under the linear and mixed mechanisms, whereas

Calibrated-SVM became increasingly competitive as the nonlinear sample size increased. The paired-regret panels show the magnitude of the performance loss when a model was not the best within a given replication.

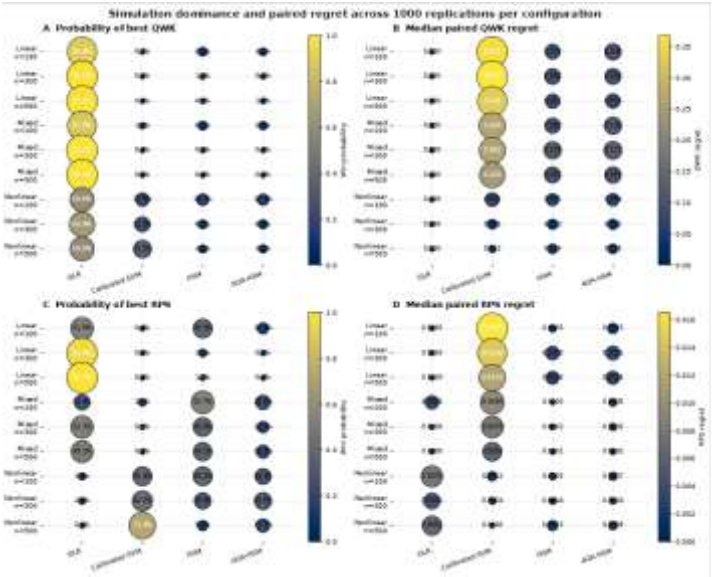

Fig 1. Simulation dominance and paired regret across the nine scenario–sample-size configurations.

The tuning behaviour was consistent with these changes in data structure. Selection of τ=0, which reduces AOA–HSM exactly to HSM, occurred in 45.20%, 67.57%, and 69.90% of the training selections under the linear mechanism for n=100, 300, and 500, respectively. The corresponding frequencies under the mixed mechanism were 29.57%, 52.47%, and 58.13%. In contrast, under the nonlinear mechanism, selection of the largest candidate value, τ=3, increased from 20.73% at n=100 to 35.77% at n=500.

A similar pattern was observed for SVM kernel selection. The RBF kernel was selected in approximately 74.0%–81.5% of the linear configurations, 82.1%–94.4% of the mixed configurations, and 87.8%–97.6% of the nonlinear configurations. The highest selection frequency, 97.57%, occurred under the nonlinear mechanism at n=500. Together, these patterns indicate that the tuning procedure generally favoured a stronger nonlinear SVM contribution as the data-generating mechanism departed further from the linear ordinal structure.

Although τ=0 was included in the candidate grid, the value selected by inner validation was not necessarily optimal on the corresponding outer-test data. Consequently, AOA–HSM was not guaranteed to outperform HSM in every replication or configuration. This behaviour reflects the directional structure of the adaptive gate: increasing disagreement can move the fusion from equal weighting toward the Calibrated-SVM component, but not toward the OLR component.

Post-selection computation was modest for all four models. Mean computation times ranged from 0.284 to 0.491 seconds for OLR and from 0.071 to 0.218 seconds for Calibrated-SVM. The corresponding times for HSM and AOA–HSM ranged from approximately 0.356 to 0.709 seconds and were essentially identical. Thus, calculation of the adaptive weight introduced negligible additional computational cost once the two base learners had been fitted. These timings refer to model fitting and prediction after hyperparameter selection and exclude the complete inner-validation search.

Fig 2 provides a complementary view of the adaptive mechanism by showing the selected τ values, the realised adaptive weights, and the relationship between base-model disagreement and the RPS difference between AOA–HSM and HSM.

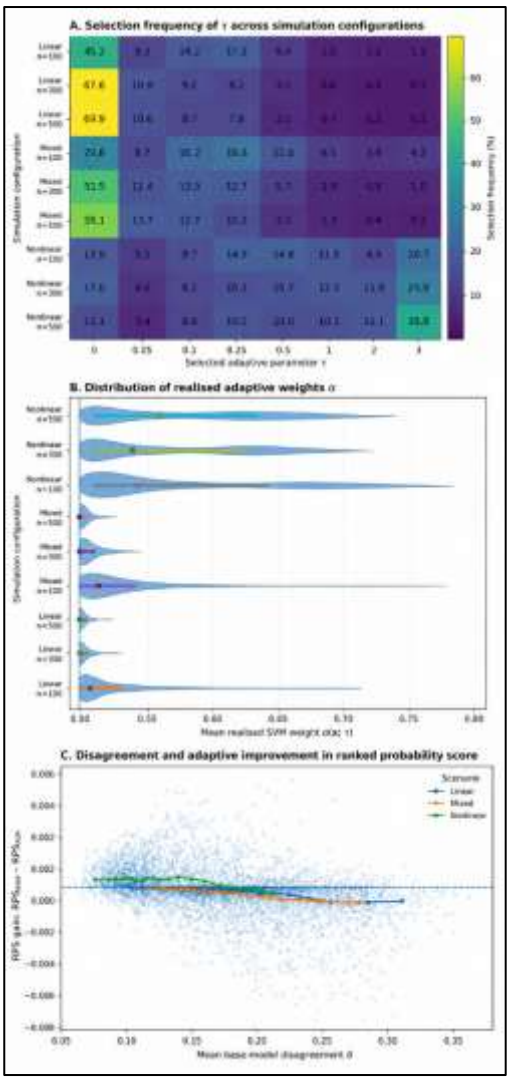

Fig 2. Adaptive-gate diagnostics for AOA–HSM.

Overall, the simulation results indicate that the relative performance of the models depended on both the underlying data structure and the evaluation criterion. OLR provided the strongest ordinal discrimination when a substantial linear component was present, HSM generally offered the more stable fusion strategy, and AOA–HSM showed its clearest benefit in probability quality under selected nonlinear conditions. The adaptive model should therefore be viewed as a context-dependent extension of fixed probability fusion rather than as a uniformly superior alternative.

4.2 Empirical Results

4.2.1 Temporal Diagnostics

The continuous CO_2 emission-growth series showed evidence of short-range temporal dependence. The lag-one autocorrelation was 0.306, and the Ljung–Box test was statistically significant at lag 1, with Q(1)=5.798 and p=0.016. Evidence of dependence weakened at longer lags: the corresponding Ljung–Box p-values were 0.050 at lag 5 and 0.110 at lag 10.

The discretised ordinal response showed weaker serial association. The lag-one Spearman correlation was ρ=0.164 with p=0.218, and the transition-based chi-square test did not reject independence between consecutive ordinal levels, with χ^2=6.468, df=4, and p=0.167. Taken together, these findings indicated that temporal dependence could not be ignored and supported the use of chronological, rather than randomly shuffled, external evaluation.

The three expanding-window test periods produced 30 genuinely out-of-sample predictions covering 1994–2023. Fig 3 shows the annual CO_2 emission-growth series, the ordinal thresholds, and the three chronologically held-out evaluation periods.

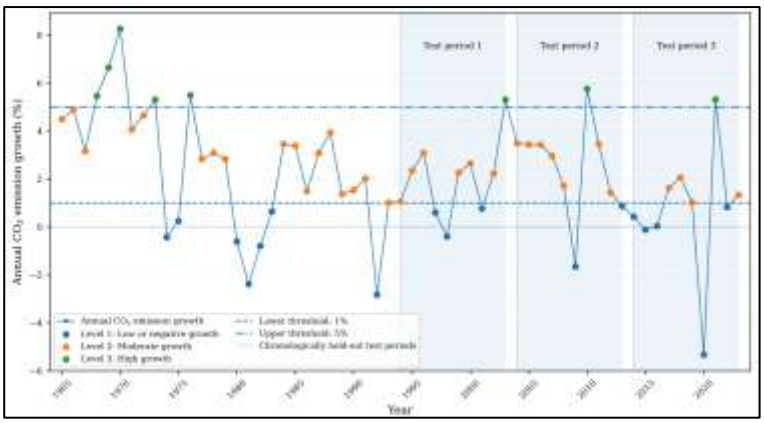

Fig3. Annual global CO_2 emission growth over 1965–2023. The horizontal lines indicate the ordinal-response thresholds of 1% and 5%. The shaded regions identify the three chronologically held-out test periods used in the expanding-window evaluation.

4.2.2 Pooled Out-of-Sample Performance

Table 5 summarises the pooled out-of-sample performance across the 30 chronologically held-out observations, together with the mean computation time per outer evaluation period.

Table 5. Pooled out-of-sample performance and computational cost.

Model	Accuracy	BACC	Macro-F1	QWK	MAE	LogLoss	Brier	Time (s)
OLR	0.600	0.463	0.420	0.303	0.433	2.063	0.681	0.398
Calibrated-SVM	0.667	0.461	0.446	0.402	0.333	0.867	0.519	1.767
HSM	0.600	0.463	0.420	0.303	0.433	0.874	0.528	2.165
AOA–HSM	0.600	0.463	0.422	0.429	0.400	0.855	0.515	3.441
Ordinal gradient boosting	0.433	0.507	0.418	0.283	0.633	3.711	0.706	21.597
Ordinal random forest	0.467	0.526	0.456	0.441	0.533	2.816	0.664	181.767

Note: BACC denotes balanced accuracy and QWK denotes quadratic weighted kappa. Larger values are preferable for Accuracy, BACC, Macro-F1, and QWK; smaller values are preferable for MAE, LogLoss, Brier score, and computation time.

Calibrated-SVM achieved the highest overall accuracy (0.667) and the lowest MAE (0.333). Its balanced accuracy was nevertheless lower than those of the two ordinal tree models, indicating that its higher exact classification accuracy was not distributed equally across the three response levels.

Ordinal random forest achieved the highest balanced accuracy (0.526), Macro-F1 (0.456), and QWK (0.441), giving it the strongest numerical performance on balanced and ordinal-sensitive classification criteria. This improvement came at a substantial computational cost, with a mean outer-period computation time of 181.767 seconds.

AOA–HSM achieved a QWK of 0.429, close to the ordinal random forest value of 0.441, while producing the lowest LogLoss (0.855) and Brier score (0.515). Its empirical advantage was therefore more apparent in ordinal agreement and probability quality than in exact classification accuracy. Relative to HSM, AOA–HSM improved QWK, MAE, LogLoss, and Brier score, although the two models had identical pooled accuracy.

4.2.3 Performance Across Temporal Test Periods

Because pooled results can mask changes across time, Table 6 provides a compact summary of accuracy, QWK, and Brier score for each held-out period. These three measures were retained to represent exact classification, ordinal agreement, and probabilistic accuracy, respectively.

Table 6. Performance across the three chronologically held-out periods. Each entry is reported as Accuracy/QWK/Brier. Larger Accuracy and QWK and smaller Brier values are preferable.

Model	1994–2003	2004–2013	2014–2023
OLR	0.700/0.348/0.459	0.600/0.103/0.693	0.500/0.359/0.890
Calibrated-SVM	0.700/0.348/0.491	0.700/0.516/0.483	0.600/0.375/0.584
HSM	0.700/0.348/0.434	0.600/0.103/0.520	0.500/0.359/0.631
AOA–HSM	0.700/0.348/0.442	0.600/0.429/0.477	0.500/0.359/0.626
Ordinal gradient boosting	0.600/0.231/0.677	0.400/0.492/0.723	0.300/0.177/0.719
Ordinal random forest	0.600/0.231/0.637	0.300/0.478/0.767	0.500/0.419/0.589

The period-specific results confirm that model performance varied over time. In 1994–2003, the four OLR/SVM-based approaches achieved the same accuracy and QWK, although HSM produced the lowest Brier score. Calibrated-SVM performed particularly well in 2004–2013, achieving an accuracy of 0.700 and a QWK of 0.516, whereas AOA–HSM produced the lowest Brier score in that period. During 2014–2023, Calibrated-SVM retained the highest accuracy, while ordinal random forest achieved the highest QWK. These changes reinforce the conclusion that no single method dominated across all periods and evaluation criteria.

4.2.4 Rare-Class Behaviour and Predictive Uncertainty

The rare Level-3 category remained the most difficult class to predict. OLR, Calibrated-SVM, HSM, and AOA–HSM did not directly classify any of the three held-out Level-3 observations as Level 3. However, AOA–HSM assigned all three cases to the adjacent Level-2 category rather than to the most distant Level-1 category. In contrast, ordinal gradient boosting and ordinal random forest each correctly classified two of the three Level-3 observations. Because only three Level-3 cases were available in the pooled test set, these class-specific findings should be interpreted descriptively.

Moving-block bootstrap analysis also indicated substantial uncertainty around several pooled estimates. For example, the 95% confidence interval for QWK was [0.270,0.545] for AOA–HSM, [0.252,0.589] for ordinal random forest, and [0.176,0.571] for Calibrated-SVM. These intervals are reported as descriptive measures of uncertainty rather than as tests of pairwise model differences. Accordingly, the small numerical differences among the leading models should not be interpreted as evidence of uniform superiority, particularly given the limited number of chronologically held-out observations.

Fig 4 summarizes the trade-off between classification performance, probability quality, and computational cost across the six models.

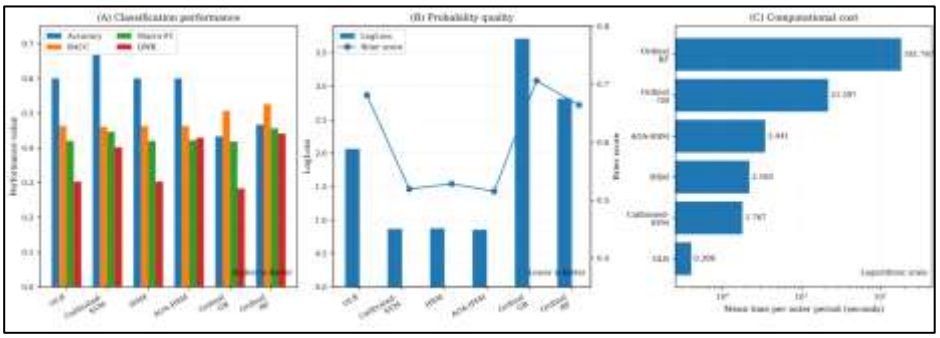

Fig 4. Comparison of pooled out-of-sample performance and computational cost.

Panel A presents accuracy, balanced accuracy, Macro-F1, and quadratic weighted kappa. Panel B presents LogLoss and Brier score, for which smaller values indicate better probability quality. Panel C presents the mean computation time per outer period on a logarithmic scale.

Overall, the empirical results show complementary strengths across the six approaches rather than uniform dominance by a single model. Calibrated-SVM provided the strongest exact classification performance, ordinal random forest performed best on several balanced and ordinal-sensitive measures, and AOA–HSM produced

the strongest probability-score results while remaining substantially less computationally demanding than ordinal random forest.

4.2.5-Model Diagnostics and Sensitivity Analyses

The proportional-odds assumption was examined using a global parallel-lines test. The test did not provide statistically significant evidence against this assumption, with $\chi^2=9.369$, df=5, and p=0.095. Thus, at the 5% significance level, there was insufficient evidence that the slope coefficients varied across the cumulative response thresholds. This result was interpreted cautiously because of the small empirical sample and the limited number of observations in Level 3.

The full-sample OLR fit was also examined descriptively. Population was positively associated with higher emission-growth levels (β=1.867, p=0.037), as was GDP growth (β=1.660, p=0.017). The strongest positive association was observed for fossil-fuel category 3 (β=4.927, p=0.005). Renewable-energy share showed a negative association that was close to conventional statistical significance (β=-1.334, p=0.058), whereas rainfall and fossil-fuel category 2 were not statistically significant. These coefficient estimates are descriptive of the full-sample association structure and were not used to assess predictive performance.

Residual diagnostics provided no evidence of substantial remaining serial dependence in the fitted OLR model. The lag-one residual autocorrelation was -0.076, and Ljung–Box tests were nonsignificant at lags 1, 5, and 10, with p-values of 0.551, 0.733, and 0.590, respectively.

Two sensitivity analyses were conducted to assess the robustness of the empirical conclusions. First, the ordinal response was reconstructed using alternative thresholds of 0.5% and 4.5%. Under this definition, ordinal random forest achieved the highest accuracy (0.733) and Macro-F1 (0.606), whereas ordinal gradient boosting produced the highest QWK (0.469). HSM and AOA–HSM each achieved an accuracy of 0.700 and a QWK of 0.416, while AOA–HSM retained the lowest LogLoss (0.754) and Brier score (0.428). These results indicate that classification-based rankings were somewhat sensitive to the choice of response thresholds, whereas the probability-score performance of AOA–HSM remained comparatively stable.

Second, hyperparameter selection was repeated using fully chronological inner validation. Under this stricter procedure, τ=0 was selected in all three outer periods, causing AOA–HSM to reduce to HSM. The resulting pooled performance of the two models was therefore identical, with accuracy 0.600, balanced accuracy 0.476, Macro-F1 0.421, QWK 0.336, MAE 0.433, LogLoss 0.977, and Brier score 0.589. Because the chronological inner-validation subsets contained very few Level-3 observations, this analysis was treated as a conservative sensitivity assessment rather than as a replacement for the primary tuning strategy.

Computational requirements also differed substantially across models. Relative to OLR, Calibrated-SVM required approximately 4.44 times more computation time, AOA–HSM approximately 8.64 times more, and ordinal random forest approximately 456 times more. Despite these relative differences, the absolute computation time remained modest for all models except ordinal random forest.

Overall, the diagnostic and sensitivity analyses did not identify a single model that dominated across all criteria. Calibrated-SVM was strongest for exact classification and MAE in the primary empirical analysis, ordinal random forest performed best on several balanced and ordinal-sensitive classification measures, and AOA–HSM provided the best probability-score performance. The results therefore support a criterion-dependent interpretation of model performance rather than a claim of universal superiority.

Discussion

The findings show that the value of combining OLR and Calibrated-SVM depends on the underlying data structure and the evaluation criterion. No model dominated across all settings, indicating that ordinal discrimination, exact classification, and probability quality should be considered as distinct objectives.

In the simulation study, OLR achieved the strongest QWK under the linear and mixed mechanisms. This result is consistent with the alignment between the cumulative-logit structure and the ordered latent response. It also confirms that increased model flexibility does not necessarily improve ordinal classification when a simpler model adequately represents the data structure. HSM generally outperformed the standalone SVM on ordinal measures and was more stable than AOA–HSM in the linear and mixed scenarios, suggesting that the OLR probabilities helped stabilise the nominal SVM predictions.

AOA–HSM showed its clearest advantage under nonlinear conditions, where it achieved the lowest RPS at n=100 and n=300. Its main benefit was therefore related to the quality of the predicted probability distribution rather than consistent improvement in QWK. The increasing selection of larger τ values and the RBF kernel under nonlinear mechanisms indicates that the tuning procedure responded to nonlinear structure. Nevertheless, the adaptive model did not uniformly outperform HSM because its disagreement-based gate is directional and can only increase the SVM contribution. It does not determine which base learner is locally more accurate, and greater disagreement may therefore increase the weight of an unreliable SVM prediction.

The empirical results also produced criterion-dependent rankings. Calibrated-SVM achieved the highest accuracy and lowest MAE, whereas ordinal random forest achieved the highest balanced accuracy, Macro-F1, and QWK. AOA–HSM achieved the lowest LogLoss and Brier score and the second-highest numerical QWK. Its principal empirical advantage was therefore probability-score performance rather than exact classification.

The ordinal tree models identified two of the three Level-3 observations, whereas the remaining models did not predict Level 3 directly. However, this apparent advantage should be interpreted cautiously because the pooled test sample contained only three Level-3 cases.

The moving-block bootstrap confidence intervals were wide and overlapping, so the numerical differences among AOA–HSM, Calibrated-SVM, and ordinal random forest do not establish conclusive superiority. The sensitivity analyses further showed that model rankings depended on the response thresholds and the inner-validation design. Under fully chronological inner validation, $\tau=0$ was selected in all periods, causing AOA–HSM to reduce to HSM. This finding suggests that the empirical benefit of adaptation becomes weaker when hyperparameter selection preserves the temporal ordering completely.

These results have two practical implications. First, ordinal problems should not be evaluated using accuracy alone, because models with similar accuracy may differ substantially in QWK, balanced accuracy, or probability scores. Second, adaptive fusion should be viewed as a context-dependent extension of fixed probability averaging rather than as a universally superior model. OLR remains attractive when interpretability, parsimony, and linear ordinal structure are central; Calibrated-SVM is useful for exact classification; ordinal tree models may improve minority-class recognition; and AOA–HSM may be preferable when probability-score performance under nonlinear heterogeneity is the main objective.

Several limitations should be acknowledged. The empirical analysis was based on only 59 annual observations, including eight Level-3 cases in the full sample and three in the pooled test periods. The analysis also used a single global time series without independent external validation. In addition, the response thresholds were operational cut-offs rather than internationally established standards, and the primary stratified inner-validation procedure did not preserve temporal ordering completely. Finally, the AOA–HSM gate does not estimate local base-model reliability and does not guarantee improvement over fixed fusion. Future research should evaluate the framework using larger country-level or panel datasets, fully chronological tuning, external validation, bidirectional gating mechanisms, and more detailed probability-calibration assessment.

Conclusion

This study developed a unified probability-fusion framework combining ordinal logistic regression and a probability-calibrated support vector machine through fixed and adaptive weighting schemes. The proposed AOA–HSM preserves valid class probabilities and reduces exactly to HSM when $\tau=0$, but it does not guarantee superior prediction in all settings.

The simulation results showed that OLR provided the strongest ordinal discrimination under linear and mixed mechanisms, whereas AOA–HSM showed its main benefit in probability-score performance under selected nonlinear conditions. In the empirical application, Calibrated-SVM achieved the highest accuracy, ordinal random forest achieved the strongest balanced ordinal performance, and AOA–HSM obtained the best LogLoss and Brier score. Therefore, no model dominated across all evaluation criteria.

Overall, the proposed adaptive framework is best viewed as a context-dependent extension of fixed probability fusion rather than a universally superior classifier. Its usefulness is greatest when nonlinear predictive information and probability quality are important, while simpler models may remain preferable when interpretability, parsimony, or ordinal discrimination is prioritised. Future research should assess the framework using larger external datasets, fully chronological tuning, and adaptive gates that estimate the local reliability of both base learners.

References

1. Agresti, A.: Analysis of Ordinal Categorical Data, 2nd edn. Wiley, Hoboken, NJ (2010). https://doi.org/10.1002/9780470594001.
2. Bala´c, N., Mileusni´c, Z., Dragiˇcevi´c, A., Milanovi´c, M., Rajkovi´c, A., Miodragovi´c, R.,
3. E´cim-Djuric, O.: Implementation of XGBoost Models for Predicting CO2 Emission and Specific Tractor Fuel Consumption. Agriculture 15(11), 1209 (2025)
4. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: Proceedings of the 20th International Conference on Pattern Recognition, pp. 3121–3124 (2010). https://doi.org/10.1109/ICPR.2010.764.
5. Brant, R.: Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics 46(4), 1171–1178 (1990) https://doi.org/10.2307/2532457
6. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998) https://doi.org/10.1023/A:1009715923555.
7. Cardoso, J.S., Costa, J.F.P.: Learning to classify ordinal data: The data replication method. Journal of Machine Learning Research 8, 1393–1429 (2007).
8. Chu, W., Keerthi, S.S.: New approaches to support vector ordinal regression. Neuro-computing 55(1–2), 23–43 (2005) https://doi.org/10.1016/j.neucom.2004.11.007.
9. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27–12727 (2011)

https://doi.org/10.1145/1961189.1961199.
10. Cohen, J.: Weighted kappa: Nominal scale agreement with provision for scaled dis-agreement or partial credit. Psychological Bulletin 70(4), 213–220 (1968) https://doi.org/10.1037/h0026256.
11. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273–297 (1995) https://doi.org/10.1007/BF00994018.
12. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple Classifier Systems, (2000). https://doi.org/10.1007/3-540-45014-9 1.
13. Epstein, E.S.: A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology 8(6), 985–987 (1969)
https://doi.org/10.1175/1520-0450(1969)008⟨0985:ASSFPF⟩2.0.CO;2
14. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200), 675–701 (1937) https://doi.org/10.1080/01621459.1937.10503522.
15. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102(477), 359–378 (2007) https: //doi.org/10.1198/016214506000001437.
16. Harrell, F.E.: Ordinal logistic regression. In: Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, pp. 311–325. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19425-7 13.
17. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284 (2009) https://doi.org/10.1109/TKDE.2008.239
18. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. Advances in Large Margin Classifiers (2000) https://doi.org/10.1023/A:1007560818575.
19. Hsu, C.-W., Lin, C.-J.: A comparison of methods for multiclass support vector. IEEE Transactions on Neural Networks 13(2), 415–425 (2002) https: //doi.org/10.1109/72.991427.
20. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2), 65–70 (1979). Intergovernmental Panel on Climate Change: Climate change 2023: Synthesis report. Technical report, IPCC, Geneva, Switzerland (2023). https://doi.org/10.59327/IPCC/AR6-9789291691647. https://www.ipcc.ch/report/ar6/syr/
21. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE.Transactions on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)
https://doi.org/10.1109/34.667881
22. Lin, H.-T., Lin, C.-J., Weng, R.C.: A note on platt's probabilistic outputs for support vector machines. Machine Learning 68(3), 267–276 (2007) https://doi.org/10.1007/s10994-007-5018-6
23. McCullagh, P.: Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological) 42(2), 109–142 (1980)
24. Pachauri, R.K., Allen, M.R., Barros, V.R., et al.: Climate Change 2014: Synthesis
Report. IPCC, Geneva, Switzerland (2014)
25. Pitawela, D., Carneiro, G., Chen, H.-T.: Cloc: Contrastive learning for ordinal classification with multi-margin n-pair loss. arXiv preprint (2025)
26. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, A.J., Bartlett, P., Sch¨olkopf, B., Schuurmans,
27. D. (eds.) Advances in Large Margin Classifiers, pp. 61–74. MIT Press, Cambridge, MA (1999)
28. Sharabiani, M.T.A., Bottle, A., Mahani, A.S.: Ogboost: A python package for ordinal gradient boosting. arXiv preprint (2025)
29. Shashua, A., Levin, A.: Ranking with large margin principle. In: Advances in Neural
Information Processing Systems (2003). https://doi.org/10.1145/944919.944941.
30. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management 45(4), 427–437 (2009) https://doi.org/10.1016/j.ipm.2009.03.002.
31. Sch¨olkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA (2002). https://doi.org/10.7551/mitpress/4175.001.0001.
32. Stern, N.: Economics: Current climate models are grossly misleading. Nature 530(7591), 407–409 (2016) https://doi.org/10.1038/530407a.
33. United Nations Framework Convention on Climate Change: The Paris Agreement. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement. Accessed: 16 Mar 2025 (2015)
34. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin 1(6),
80–83 (1945) https://doi.org/10.2307/3001968.