Variance-Aware Penalized Panel Models for Temporal Risk Detection from Wearable Sensor Data

Zihao Wang[1] and Min Lu[1*]

[1*]Division of Biostatistics, Miller School of Medicine, University of Miami, Miami, FL, 33136, USA.

*Corresponding author(s). E-mail(s): m.lu6@umiami.edu;

Abstract

Background: Wearable devices generate continuous, high-resolution physiological data that offer opportunities for real-time assessment of stress, arousal, and early physiological deterioration, but existing pipelines often treat variability as nuisance noise or rely on labeled classifiers. We present a computational approach for subject-adaptive temporal risk detection that explicitly separates conditional mean and variance dynamics in high-dimensional multisubject sensor data.

Results: The proposed penalized panel ARX–GARCHX model integrates subject-specific baselines, shared autoregressive dynamics, sparse multimodal covariate effects, and covariate-dependent volatility. It produces an exceedance-based risk score that estimates the conditional probability of crossing an individualized physiological threshold. Simulation experiments across stable, seasonal, transient-regime, and sustained-regime settings showed that modeling covariate-driven variance improves recovery of threshold-exceedance risk when volatility is structured. In the Wearable Stress and Affect Detection (WESAD) demonstration, the score provided an interpretable, label-free temporal summary that separated stress-associated windows more clearly than raw heart-rate summaries and remained lightweight for streaming use.

Conclusions: Variance-aware penalized panel modeling provides a reproducible methodology for converting noisy wearable streams into subject-adaptive risk-state scores. It is intended for translational feature extraction, with prospective validation required before clinical decision support.

Keywords: wearable sensors, digital biomarkers, temporal risk detection, heteroscedasticity, panel time series, GARCHX, physiological monitoring

1 Background

Wearable sensing systems generate continuous multisubject, multimodal physiological streams, including heart rate, electrocardiography (ECG), electrodermal activity (EDA), respiration, body temperature, electromyography (EMG), and acceleration. These signals are increasingly used to study stress, autonomic regulation, sleep-wake patterns, activity, and broader psychophysiological states in settings that are closer to daily life than traditional clinic-based assessment [1–4]. For translational medicine, such data may provide scalable, low-burden measurements that complement clinical assessments and patient-reported outcomes.

A major obstacle is that wearable physiological signals are not simply noisy observations of stable latent states. Their temporal structure is affected by motion artifacts, device placement, context, individual baseline differences, and heteroscedastic variability. Much recent work has focused on supervised machine-learning and deep-learning models for wearable affect recognition and physiological state detection [5–10]. Related signal-processing work has emphasized denoising, adaptive filtering, instantaneous-frequency estimation, forecasting, and anomaly detection for wearable or biomedical sensor streams [11–17]. These approaches demonstrate the value of continuous physiological monitoring, but many rely on dense labels, short-window classification, or feature extraction without explicitly separating changes in location from changes in volatility. Distinguishing changes in location from changes in volatility is important in biomedical and translational studies, where a risk-relevant physiological deviation may depend on a participant's own baseline and variance pattern rather than on a fixed population threshold [18–20].

Model-based temporal analysis provides a complementary route. Autoregressive models with exogenous covariates (ARX) are useful for dynamic input-output relationships [21], and generalized autoregressive conditional heteroscedasticity models with covariates (GARCHX) allow the conditional variance to depend on lagged variation and external predictors [22–26]. However, standard implementations are usually designed for single series and low-dimensional inputs. Wearable studies, in contrast, often involve repeated measurements from multiple participants and many correlated multimodal features. A translationally useful approach should therefore pool information across participants, retain subject-specific baselines, select informative covariates, and return a risk score that can be interpreted on a probability scale. Dynamic panel modeling [27] and sparse penalization [28–31] provide useful building blocks, but they are rarely integrated with covariate-driven volatility modeling for wearable physiological risk detection. Related work on variance-guided regression for heteroscedastic data has shown that explicitly leveraging conditional variance structure can improve prediction [32]. In many sensing applications, the goal extends beyond forecasting to deriving a *temporal risk score* that captures mean dynamics and variance fluctuations under realistic noise. Wearable streams often show sharp volatility changes driven by motion or context, motivating explicit variance modeling when extracting stable temporal representations from noisy signals.

We propose a variance-aware penalized panel model for temporal risk detection from wearable sensor data. The model combines a panel ARX mean equation with a

GARCHX variance equation, uses sparsity penalties to handle high-dimensional multimodal covariates, and outputs an exceedance-based risk score. The score estimates the conditional probability that a target physiological signal exceeds a subject-specific threshold after accounting for recent history, covariates, and time-varying variance. Figure 1 summarizes the workflow. Panel (A) shows how multimodal wearable streams are processed through the penalized ARX–GARCHX framework to obtain variance-aware temporal representations, while Panel (B) illustrates pronounced volatility changes in windowed heart-rate data from the Wearable Stress and Affect Detection (WESAD) dataset [4]; here `HR_mean` denotes the mean heart rate (in beats per minute) within each 60-s window, the target signal used throughout.

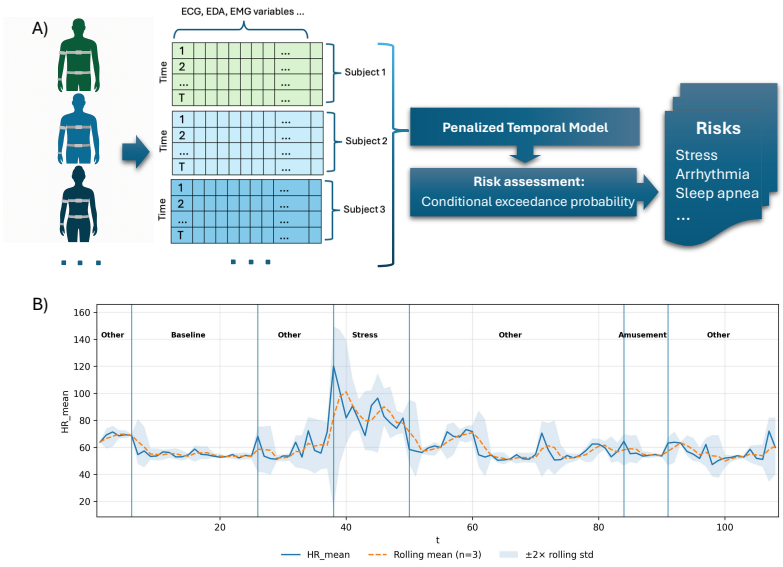


Fig. 1: Conceptual workflow for temporal risk detection from wearable sensor data. Multisubject physiological streams are represented as window-level features, analyzed with a penalized temporal model, and summarized through a conditional exceedance probability that can be used as a subject-adaptive risk assessment. Any downstream conditions shown in Panel (A), such as arrhythmia or sleep apnea, are *hypothetical* illustrative applications and are not analyzed in this study; the only empirical demonstration here is the WESAD stress benchmark. The lower panel illustrates heart-rate dynamics across annotated WESAD segments.

The aims of this study were: (i) to develop a penalized panel ARX–GARCHX model for high-dimensional multisubject wearable streams; (ii) to define an interpretable exceedance-based risk score; (iii) to evaluate recovery of the risk score in simulations with structured volatility; and (iv) to demonstrate proof-of-concept performance for stress-associated temporal risk detection using the WESAD benchmark dataset [4, 33].

2 Methods

2.1 Penalized panel ARX–GARCHX model

Let $s = 1, \ldots, S$ index subjects and $t = 1, \ldots, T$ index time. All subjects share the same time length T, so we consider a balanced panel. We observe $\{(y_{s,t}, x_{s,t})\}$, where $y_{s,t}$ is a univariate sensor-derived signal of interest (e.g., heart rate), and $x_{s,t} \in \mathbb{R}^d$ denotes a high-dimensional vector of multimodal covariates (e.g., engineered features or concurrent sensor channels). For the response variable $y_{s,t}$, we consider an autoregressive panel data model [27] of the form

$$y_{s,t} = \alpha_s + \sum_{i=1}^{P} \theta_i \, y_{s,t-i} + \sum_{j=0}^{Q} x_{s,t-j}^{\top} \beta_j + \sigma_{s,t} \, \varepsilon_{s,t}, \tag{1}$$

where $\mathbb{E}[\varepsilon_{s,t} \mid x_{s,t}, \mathcal{F}_{t-1}] = 0$, $\mathrm{Var}(\varepsilon_{s,t} \mid x_{s,t}, \mathcal{F}_{t-1}) = 1$, and $\mathcal{F}_{t-1}$ denotes the information set generated by all past observations and covariates up to time $t-1$. The innovations $\varepsilon_{s,t}$ thus have zero conditional mean and unit conditional variance given the past, making $\sigma_{s,t}$ the conditional standard deviation. Equivalently, define $\mu_{s,t} := \alpha_s + \sum_{i=1}^{P} \theta_i \, y_{s,t-i} + \sum_{j=0}^{Q} x_{s,t-j}^{\top} \beta_j$ so that $y_{s,t} = \mu_{s,t} + \sigma_{s,t} \varepsilon_{s,t}$.

The conditional variance model is

$$\sigma_{s,t}^2 = \omega_s + \sum_{\ell=1}^{L} b_\ell \sigma_{s,t-\ell}^2 + \sum_{r=1}^{R} a_r e_{s,t-r}^2 + \sum_{m=0}^{M} x_{s,t-m}^{\top} \gamma_m, \tag{2}$$

where $e_{s,t} = y_{s,t} - \mu_{s,t}$ and ω_s is a subject-specific baseline variance. To ensure $\sigma_{s,t}^2 > 0$, we (i) enforce $a_r, b_\ell \geq 0$, $\omega_s \geq \omega_{\min} > 0$, and (ii) project the linear predictor to a positive set when needed. Implementation safeguards are detailed in Supplementary Section S2.

Sharing vs. individual effects.

Intercepts α_s (mean) and ω_s (baseline variance) are *subject-specific*, while $\{\theta_i\}, \{\beta_j\}, \{\gamma_m\}, \{a_r\}, \{b_\ell\}$ are *shared* population parameters.

2.2 Penalized quasi-likelihood

Parameters are estimated by penalized Gaussian quasi-likelihood [34]:

$$\min_{\alpha, \theta, \beta, \omega, a, b, \gamma} \sum_{s=1}^{S} \sum_{t=1}^{T} \left\{ \log \sigma_{s,t}^2 + \frac{(y_{s,t} - \mu_{s,t})^2}{\sigma_{s,t}^2} \right\} + P_\lambda(\beta) + P_\rho(\gamma), \tag{3}$$

where P_λ and P_ρ may be the ℓ_1, elastic net, smoothly clipped absolute deviation (SCAD) [35], minimax concave penalty (MCP) [36], or group penalties. In this paper, we employ the ℓ_1 penalty to select informative covariates in both the conditional mean and conditional variance equations. Model fitting uses alternating updates of the mean and variance blocks.

Although the focus of this work is methodological and applied, the proposed penalized estimator enjoys standard high-dimensional guarantees. Under regularity, restricted-strong-convexity, and score-bound conditions on the panel process (Assumptions S1–S3 in the Supplementary Material), the penalized ARX–GARCHX estimator $(\widehat{\beta}, \widehat{\gamma})$ attains an excess-risk rate of order $(s_\beta + s_\gamma) \log p/(ST)$, where (s_β, s_γ) are the sparsity levels of the oracle pair $(\beta^\star, \gamma^\star)$ and p is the ambient dimension; the formal statement and full proof are given in Supplementary Section S1 (Theorem S1).

2.3 Algorithm: alternating penalized estimation

Given current conditional variance estimates, the subject-specific intercepts in the conditional mean equation are updated by weighted least squares, and the shared mean coefficients (θ, β) are obtained from a weighted ℓ_1-penalized regression. The variance block uses a decoupled update: (ω, a, b) is refit by constrained nonlinear optimization with γ held fixed, after which γ is updated by a proximal-gradient step under positivity safeguards. A forward recursion then refreshes $\sigma^2_{s,t}$ for the next outer iteration. The complete pseudocode (Algorithm S1), positivity enforcement (Remark on the variance recursion), and convergence/tuning details are given in Supplementary Section S2. For all experiments, covariates were standardized within each subject and the lag orders were fixed to $(P, Q, L, R, M) = (1, 0, 1, 1, 0)$, reflecting the short-memory structure typical of wearable sensor data; penalties (λ, ρ) were selected by blockwise time-series cross-validation that preserves temporal ordering within each subject.

2.4 Exceedance-based risk score

The fitted model provides estimated conditional mean $\widehat{\mu}_{s,t}$ and conditional standard deviation $\widehat{\sigma}_{s,t}$. For a subject-specific threshold c_s, we define the temporal risk score as

$$\pi_{s,t}(c_s) = \Pr(Y_{s,t} > c_s \mid \mathcal{F}_{t-1}, x_{s,t}) = 1 - F_\varepsilon \left(\frac{c_s - \mu_{s,t}}{\sigma_{s,t}} \right), \tag{4}$$

where F_ε is the innovation distribution function. We used the Gaussian distribution for the primary analysis; a standardized Student-t distribution can be used for heavier-tailed signals. Choosing c_s as a participant-specific quantile of the observed signal makes the resulting score comparable across participants while preserving subject-specific baselines.

The score is not a diagnosis. It is an interpretable physiological risk-state indicator: high values indicate that, given the recent history and multimodal covariates, the target signal is likely to exceed an individualized threshold. The score and its interpretation are explicitly conditional on the choice of threshold c_s: $\pi_{s,t}(c_s)$ answers "how likely is the signal to exceed c_s right now," so a different c_s defines a different exceedance event and a different score. The threshold should therefore be chosen to match the intended use—for example a within-subject quantile for baseline-relative monitoring, or a clinically anchored value when one is available—and reported alongside the score. We return to this dependence, and to how c_s is set in practice, in Sections 4.2 and 5.

3 Simulation Studies

3.1 Data generating scenarios

We evaluated the proposed variance-aware autoregressive model on synthetic panel datasets designed to challenge both mean tracking and variance modeling under smooth and piecewise changes. For subjects $s = 1, \ldots, S$ and times $t = 1, \ldots, T$, we generated innovations $\varepsilon_{s,t}$ that were conditionally independent over time given $\mathcal{F}_{t-1}$ and independent across subjects. We considered either Gaussian or standardized Student-t innovations,

$$\varepsilon_{s,t} \sim \begin{cases} N(0,1), & \text{Gaussian}, \\ t_\nu / \sqrt{\nu/(\nu-2)}, & \text{Student-}t\ (\nu > 2), \end{cases} \qquad \mathrm{Var}(\varepsilon_{s,t}) = 1. \tag{5}$$

The exogenous vector $X_{s,t}$ combined a low-dimensional signal component, a binary regime indicator that induced heteroscedasticity, and a high-dimensional block of nuisance noise:

$$X_{s,t} = \left(Z_{s,t}^\top, X_{s,t}^{\mathrm{bin}}, W_{s,t}^\top\right)^\top, \quad Z_{s,t} \in \mathbb{R}^2, \quad X_{s,t}^{\mathrm{bin}} \in \{0,1\}, \quad W_{s,t} \in \mathbb{R}^{d_{\mathrm{noise}}}. \tag{6}$$

The two-dimensional signal block $Z_{s,t}$ carried the true signal. As shown in Table 1, we considered independent bivariate standard normal draws, coordinate-wise autoregressive processes with mild persistence, and a seasonal sine-cosine pair. These generators represent independent baseline variation, short-memory persistence typical of sensor drift, and diurnal or circadian-like variation. The binary regime indicator $X_{s,t}^{\mathrm{bin}}$ modulated volatility and was generated either by a within-subject median trigger or by piecewise-constant segments. The median trigger captures personalized volatility bursts, whereas the piecewise design creates sustained regime shifts resembling activity or context changes. The nuisance block $W_{s,t}$ followed a weakly persistent autoregressive process,

$$W_{s,t}^{(\ell)} = \rho W_{s,t-1}^{(\ell)} + \xi_{s,t}^{(\ell)}, \qquad \xi_{s,t}^{(\ell)} \overset{i.i.d.}{\sim} N(0,1), \qquad W_{s,0}^{(\ell)} \sim N\left(0, \frac{1}{1-\rho^2}\right), \tag{7}$$

for $\ell = 1, \ldots, d_{\mathrm{noise}}$. Modeling nuisance covariates as autocorrelated rather than white noise makes separation of true signals from background variation more realistic.

As shown in Figure 2, in Scenario S1 the ribbon varies only modestly and $y_{s,t}$ tracks $\mu_{s,t}$ closely, indicating relatively stable variance without regime-driven shifts. In Scenario S2, sinusoidal covariates generate a smooth diurnal modulation of $\mu_{s,t}$ while ribbon width remains approximately constant, so variability is primarily mean-driven. Scenario S3 has numerous short shaded intervals that coincide with brief increases in $\sigma_{s,t}$: the ribbon widens but $\mu_{s,t}$ remains level, demonstrating transient and variance-driven episodes. In Scenario S4, extended shaded blocks coincide with sustained increases in $\sigma_{s,t}$ and more frequent large deviations, consistent with heavy-tailed innovations.

Table 1: Simulation design for Scenarios S1–S4. All scenarios share $\alpha_s \sim \mathcal{N}(0, 0.5^2)$, and the noise generator W follows the AR(1) process in Equation (7).

	S1	**S2**	**S3**	**S4**
ω_s^*	0.10	0.10	0.01	0.01
θ_i	0.40	0.50	0.40	0.40
(a_r, b_ℓ)	(0.05, 0.50)	(0.06, 0.50)	(0.06, 0.20)	(0.06, 0.20)
β_j on Z	(0.40, 0)	(0.80, 0.60)	(0.20, 0.10)	(0.80, 0.60)
γ_m on Z or X^{bin}	Z: (0.20, 0)	Z: (0.10, 0)	X^{bin}: 0.40	X^{bin}: 0.50
Innovations $\varepsilon_{s,t}$	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	t_ν (ν=8)
Signal Z	i.i.d. $\mathcal{N}_2(0, I_2)$	seasonal $(\sin, \cos)^{\S}$	AR(1)	seasonal $(\sin, \cos)^{\S}$
Regime X^{bin}	none	none	subject median†	piecewise‡
Burn-in	200	220	180	180

Note. Parameter values are held constant across indices i, r, ℓ, j, and m within each scenario. The burn-in period is used only during data generation to remove initialization transients and is discarded before model fitting. Consequently, it does not affect the estimation procedure, the estimands, or the reported Bias and RMSE. The modest differences in burn-in lengths across scenarios were chosen for simulation convenience to ensure that the retained observations are approximately from the stationary regime.

$*$ ω_s is set to be the same for all subjects s.

§ $Z^{(1)} = \sin(2\pi t/T_P)$ and $Z^{(2)} = \cos(2\pi t/T_P)$ are seasonal covariates shown in Figure 2(B,D).

† Subject-median trigger $X_{s,t}^{\text{bin}} = \mathbb{1}\{Z_{s,t}^{(1)} > \text{median}_t(Z_{s,t}^{(1)})\}$, computed per subject over the post-burn-in window; this produces frequent, short high-variance bursts.

‡ Piecewise-constant $X_{s,t}^{\text{bin}}$ alternates between 0 and 1, with run lengths jittered around a target scale T_L shared across subjects, yielding sparse, long-duration blocks.

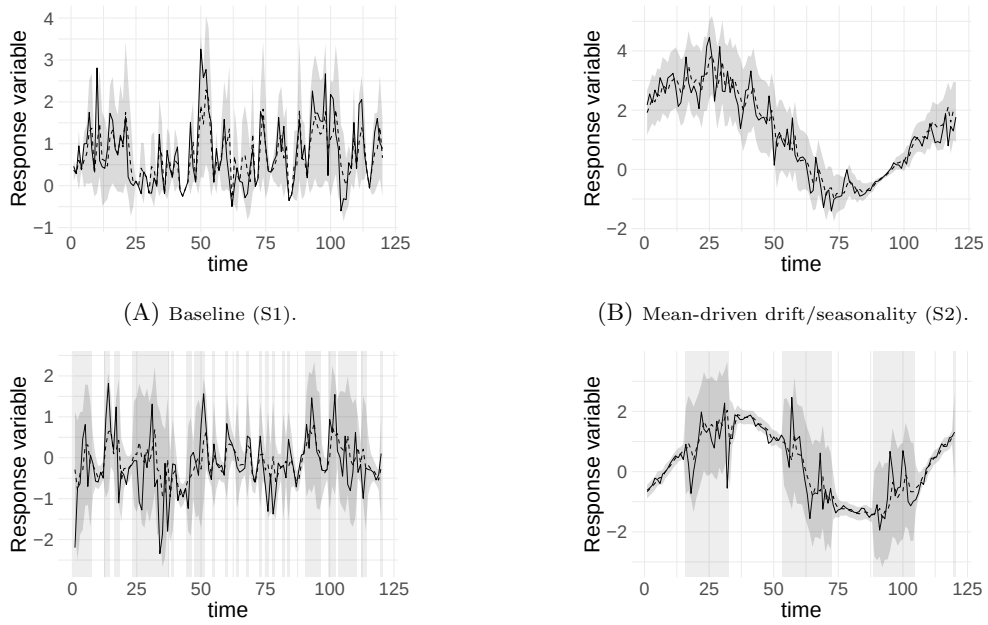
3.2 Competing models and evaluation metrics

We compare six variants within the ARX–GARCH family (Models A–F). Models A–D include covariates in the conditional mean and differ in whether the mean or variance components are penalized and whether X also enters the variance equation; Model E adds the intermediate specification with a penalized mean and an *unpenalized* X-dependent variance, and Model F is a covariate-free AR–GARCH baseline:

- **Model A**: ARX–GARCH (variance independent of X),
- **Model B**: ARX$_p$–GARCH (penalized mean; variance independent of X),
- **Model C**: ARX–GARCHX (variance includes X, unpenalized),
- **Model D**: ARX$_p$–GARCHX$_p$ (penalized mean and penalized variance),
- **Model E**: ARX$_p$–GARCHX (penalized mean; unpenalized X-dependent variance),
- **Model F**: AR–GARCH (no covariates in the mean or variance; baseline).

Model E isolates the effect of penalizing the variance covariates (E versus D) while holding the penalized mean fixed, and Model F quantifies the value added by covariates over a purely autoregressive, covariate-free specification. Here ARX$_p$ and GARCHX$_p$ denote the ARX mean and GARCHX variance equations with ℓ_1 penalties applied to the covariate coefficients β and γ, respectively.

To focus on relevant tails, we evaluate cutoffs c at pooled empirical quantiles $p \in \{0.60, 0.70, 0.80, 0.90, 0.95\}$, providing global thresholds that preserve between-subject heterogeneity at small T. We fix lag orders $(P, Q, L, R, M) = (1, 0, 1, 1, 0)$ with



(A) Baseline (S1). (B) Mean-driven drift/seasonality (S2).

(C) Regime contrast (S3). (D) Seasonal + piecewise regime (S4).

Fig. 2: Illustration of one example subject for each scenario (S1–S4) with $y_{s,t}$ (solid), $\mu_{s,t}$ (dashed), and the band $\mu_{s,t} \pm 2\sigma_{s,t}$. For Scenarios 3 and 4, gray shading denotes intervals where the regime indicator is active (X^{bin}=1).

standardized predictors. From each fitted model, we obtain $(\widehat{\mu}_{s,t}, \widehat{\sigma}_{s,t})$ and compute $\widehat{\pi}_{s,t}(c)$. For a given threshold c, we compare the estimate $\widehat{\pi}_{s,t}(c)$ with the oracle $\pi_{s,t}(c)$ via

$$\text{Bias}(c) = \frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T}\big(\widehat{\pi}_{s,t}(c) - \pi_{s,t}(c)\big), \text{ and } \text{RMSE}(c) = \sqrt{\frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T}\big(\widehat{\pi}_{s,t}(c) - \pi_{s,t}(c)\big)^2}.$$

We averaged the bias and RMSE across the five cutoff values to summarize each replication. Simulations were conducted with $S = 15$ subjects, $d_{\text{noise}} = 100$ noise covariates, and time horizons $T \in \{120, 240\}$, chosen to match the scale of our real-data application to the WESAD study. We set $\rho = 0.5$ for the AR(1) noise variables W and $\rho = 0.6$ for the signal variable Z in Scenario 3. In Scenarios 2 and 4, we used a seasonal period $T_P = T/3$ for the covariate Z, and Scenario 4 additionally used a piecewise regime run length of $T_L = 20$ for the binary covariate X^{bin}. Each scenario was replicated 100 times.

3.3 Simulation results

Across all simulated settings, the accuracy of the exceedance-based risk score is governed by whether the conditional variance is allowed to depend on covariates and whether that dependence matches the structure of the underlying volatility. When volatility is stable (Scenario 1), all six models perform similarly, so modeling covariate-driven variance costs little where it is unnecessary. Once volatility becomes structured—through seasonality (Scenario 2), regime changes (Scenario 3), or the two combined with heavy-tailed innovations (Scenario 4)—the variance-aware specifications separate clearly from the variance-agnostic baselines, attaining lower RMSE and near-zero bias in the risk score (Figure 3). The proposed specification **D** is strongest in the seasonal, wearable-like regime (Scenario 2), and its margin widens as the regime structure intensifies (Scenarios 3–4) and as the horizon lengthens to T=240; even under the heavy-tailed innovations of Scenario 4, **C** and **D** remain robust and continue to outperform the variance-agnostic baselines. Extending the comparison to all six models refines rather than overturns this picture. Model **E** (ARX$_p$–GARCHX; penalized mean, unpenalized X-dependent variance) is essentially indistinguishable from **D**—indeed the most accurate model at the longer horizon T=240, with the lowest RMSE in all four scenarios—indicating that penalizing the variance covariates chiefly stabilizes estimation at small T rather than shifting the central tendency. At the other extreme, the covariate-free baseline **F** (AR–GARCH) has by far the highest RMSE in every scenario (roughly 0.10–0.16 versus 0.05–0.09 for the covariate-driven models), quantifying the accuracy lost by ignoring exogenous covariates altogether. Taken together, the six models bracket the value of covariate-driven variance: decisive when volatility is structured, negligible when it is not, and best realized by the penalized specification that couples covariate information with regularized estimation. Full six-model results are reported in Supplementary Table S1.

Both variance-aware models **C** and **D** exhibit a mild positive bias, reflecting slight overestimation of exceedance probabilities, but the mechanism differs between them. In Model **D**, which applies an ℓ_1 penalty to the variance-covariate coefficients γ, the bias reflects shrinkage of γ toward zero, which mildly attenuates the fitted conditional variance and shifts exceedance probabilities upward through the curvature of $z \mapsto 1 - \Phi(z)$. Model **C**, by contrast, applies *no* penalty in the variance equation; its bias instead arises from finite-sample estimation noise when the high-dimensional variance-covariate vector is fit without regularization (many near-zero coefficients estimated from short series), together with the positive-variance safeguard, which together inflate small fitted variances. In both cases the effect is small. Even so, their RMSE remains lowest overall, indicating that the variance-aware formulations achieve a favorable bias–variance tradeoff and provide the most reliable exceedance-based temporal summaries across scenarios and time horizons. Replication-level Bias and RMSE means with standard errors are tabulated in Supplementary Table S1.

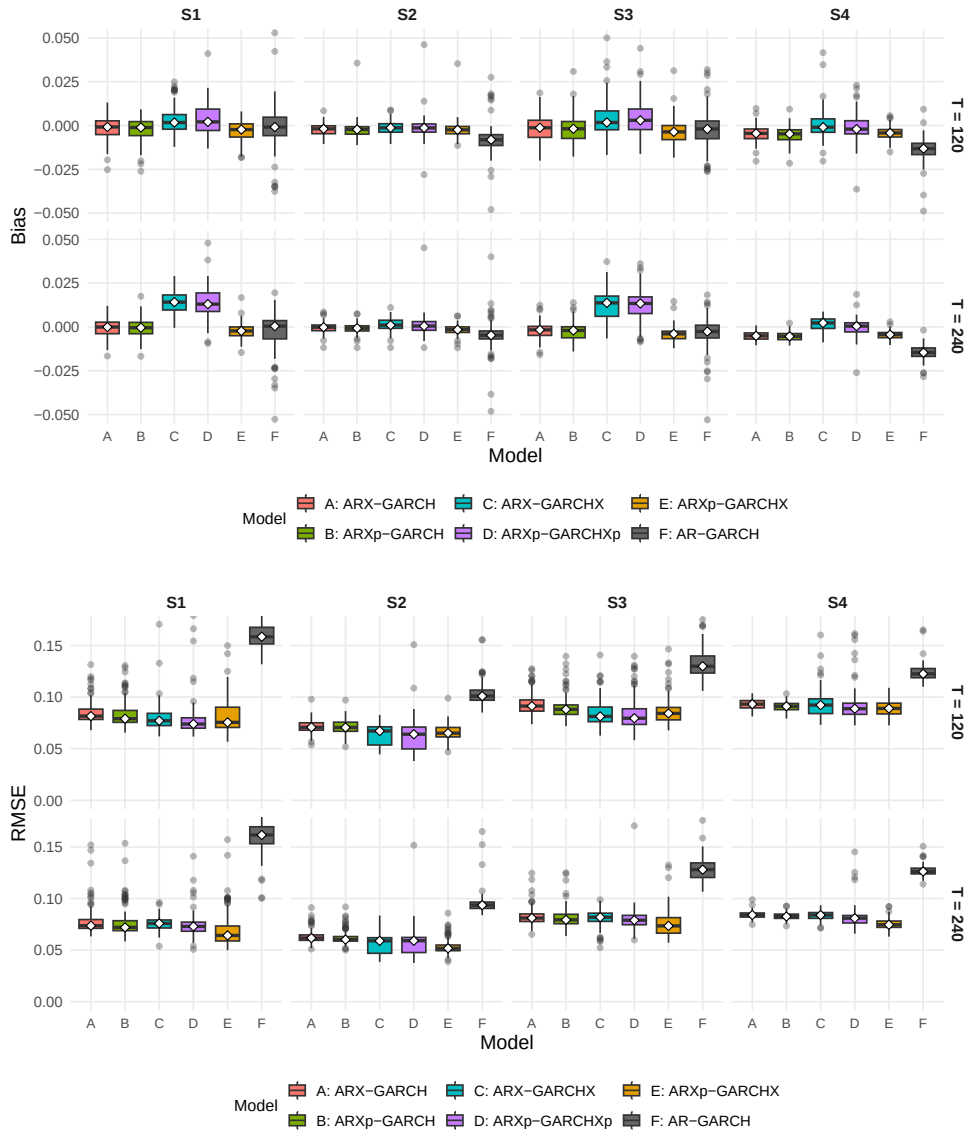


Fig. 3: Bias (top) and RMSE (bottom) of the estimated risk score across simulation Scenarios S1–S4 and Models A–F, at horizons $T = 120$ and $T = 240$ (boxplots over 100 replications; the white diamond marks the median; the legend identifies the six model variants). Model A: ARX–GARCH; Model B: ARX$_p$–GARCH; Model C: ARX–GARCHX; Model D: ARX$_p$–GARCHX$_p$; Model E: ARX$_p$–GARCHX; Model F: AR–GARCH (covariate-free baseline). The covariate-free baseline (F) has by far the highest RMSE in every setting, and the variance-aware models (C, D, E) are the most accurate, with E and D essentially indistinguishable.

4 The WESAD Lab Study

We evaluate the proposed framework using the WESAD dataset [4, 33], a controlled multimodal wearable-sensing benchmark involving 15 adult participants. Each subject completed a scripted session consisting of several annotated segments (e.g., resting, task-engaged, and amusement periods), providing clear temporal structure for analysis. Participants wore two devices—a chest-worn RespiBAN and a wrist-worn Empatica E4—so all physiological streams are time-synchronized across modalities.

The chest device provides ECG, EDA, EMG, respiration, skin temperature, and 3-axis acceleration, while the wrist device provides photoplethysmography (PPG/BVP), EDA, temperature, and acceleration. Because the protocol follows a predefined sequence of segments and the recordings are concurrent across modalities, WESAD offers clean annotations and consistent multisensor coverage suitable for multisubject or subject-wise analysis.

In our study, we focus on the chest-worn streams, as they provide the most precise heart-rate estimation. Heart rate serves as the target signal for temporal modeling due to its continuity, stability under windowing, and sensitivity to motion and other factors that drive heteroscedastic behavior in wearable data. The remaining modalities (EDA, EMG, RESP, ACC, TEMP) are incorporated as covariates to inform the covariate-driven mean and variance dynamics within the proposed variance-aware temporal risk model. The dataset annotations are used only for visualization and reference-based evaluation, not for model fitting.

4.1 Data preparation

Preprocessing and feature extraction were performed in Python using toolboxes designed for physiological-signal analysis. NeuroKit2 was used for ECG/PPG peak detection and for deriving HR, heart-rate variability (HRV), respiration, and EDA features [37]. For chest-worn devices, all streams were segmented into fixed 60-s windows with a 60-s hop. A window was retained only if at least 80% of its samples were finite across the required channels. To ensure temporal alignment across modalities, the raw streams were anti-aliased and resampled onto a common low-rate timeline prior to computing windowed features and cross-modal correlations.

Feature construction was carried out per 60-s window. From the synchronized multimodal series, we computed (i) per-series moments (mean, standard deviation, skewness, and excess kurtosis) and (ii) all pairwise Pearson cross-modal correlations to capture coordinated multimodal changes, a common pattern in wearable sensing where channels co-modulate under shared motion or environmental drivers. This produced a larger set of candidate window-level features, which were reduced to $d = 77$ by removing one member of each pair of highly collinear features (absolute Pearson correlation above 0.90, using `caret::findCorrelation`); the retained 77 features enter the model and are the covariates listed in Table 2. Windows with fewer than 80% finite samples in a required channel were dropped before feature construction, and any window with a non-finite feature value after construction was excluded; no imputation was used. After these steps, 1,427 windows remained across the 15 participants (87–117 windows per participant; participants S1 and S12 are excluded following the

WESAD convention). The complete list of the 77 features is given in Supplementary Section S8 (Supplementary Table S6), and the retained window count for each participant is given in Supplementary Table S3. Because participants contribute different numbers of windows, the analyzed panel is in fact *unbalanced*; the balanced-panel notation of Section 2.1 is used only for expositional clarity, and the estimator constructs all lags within each participant separately (see Supplementary Section S2), so unequal series lengths are handled without truncation or padding.

For reference-based evaluation, each window was also assigned a dataset annotation (baseline, stress, amusement, or transition) using majority vote over the WESAD labels. These annotations were not used when estimating the variance-aware panel model but served only for visualization and for assessing whether the derived exceedance-based risk score provides clearer temporal differentiation than the raw heart-rate values.

Table 2: Chest-worn multimodal features used in the models (60 s windows). Summaries are grouped by sensing block.

Block	**Summary of features**
ECG / HRV	Short-term variability (RMSSD/pNNx), LF/HF power and ratio, simple Poincare geometry, and compact entropy/complexity indices
Respiration & BRV	Breathing rate and amplitude (levels and variability), inspiratory/expiratory durations, duty cycle, and breath-to-breath variability
EDA	Tonic level (mean/dispersion/trend) and phasic activity (SCR count, typical amplitude, cumulative phasic area)
EMG	Broadband power, median frequency, spectral entropy, and zero-crossing rate
Accelerometry	Movement intensity (RMS) and dynamics (jerk), axis medians, high-activity tail (P90), plus overall SD/skewness/kurtosis
Skin temperature	Level, per-minute slope (trend), and detrended variability
Cross-channel correlations	Short-lag correlations among RESP, EDA, ACC, TEMP, and EMG to capture coordinated multimodal changes

Note. RMSSD: root mean square of successive differences; pNNx: percentage of successive NN intervals differing by more than x ms; LF/HF: low-/high-frequency power ratio; BRV: breathing rate variability; SCR: skin conductance response; RMS: root mean square; SD: standard deviation; RESP: respiration; ACC: accelerometer; TEMP: skin temperature.

4.2 Model specification and analysis plan

We model the 60-s windowed heart-rate mean (`HR_mean`), denoted $y_{s,t}$, as a stable, continuous signal that captures temporal fluctuations in wearable data. An ARX–GARCHX model with subject-specific intercepts and contemporaneous covariates in both the mean and variance components is fitted according to Equations (1) and (2). The lag orders are fixed at $(P, Q, L, R, M) = (1, 0, 1, 1, 0)$, a structure suitable for short-memory wearable streams, and ℓ_1 regularization is applied to (β, γ) to promote parsimony and mitigate collinearity among the multimodal features.

To provide a normalized exceedance-based risk score across subjects, we evaluate the conditional exceedance probability (4) at subject-specific thresholds. For a chosen quantile p, we set $c_s = c_s(p) := \text{Quantile}_p(\{y_{s,t}\}_t)$ – the p-th quantile of subject s's observed signal – and substitute this c_s into (4) under the assumed innovation distribution (Gaussian or standardized t). For the WESAD analysis we set $p = 0.70$, so that $c_s(0.70)$ is the 70th percentile of each participant's observed `HR_mean` series; this individualized threshold makes the score comparable across participants while preserving subject-specific baselines. The resulting exceedance probabilities summarize the joint mean–variance dynamics in a compact, interpretable form that is independent of downstream labels.

To clarify the interpretation and estimation of the proposed score, $\widehat{\pi}_{s,t}(c_s)$ denotes the estimated conditional probability that `HR_mean` exceeds the individualized threshold c_s; it is *not* a probability of stress or of any clinical event. Because the conditional variance model uses contemporaneous covariates $x_{s,t}$ (lag order M=0), $\widehat{\pi}_{s,t}$ represents the current window rather than a one-step-ahead forecast. A forecasting variant can be obtained by restricting both the mean and variance equations to strictly lagged covariates. In this demonstration, the threshold c_s, the within-subject covariate standardization, and the subject-specific parameters (α_s, ω_s) are all derived from the analyzed recordings. In a prospective deployment, these quantities would instead be obtained during an initial subject-specific calibration period and then held fixed during subsequent monitoring. Specifically, $c_s(0.70)$ is defined as the 70th percentile of each subject's `HR_mean` distribution over the full recording. It is therefore computed without using any stress annotations, although it is estimated from the entire recording, including the windows that are later used for evaluation. A sensitivity analysis that instead estimates c_s from a leak-free calibration period is reported in Supplementary Section S6 and summarized in Section 4.3.

Evaluation endpoint.

For reference-based evaluation, we define a binary window-level endpoint. The positive class consists of windows annotated as *stress* (WESAD label 2), whereas the negative class pools all other retained windows (baseline, amusement, and transition/other). Transition and amusement windows are therefore treated as non-stress rather than discarded. After preprocessing, 1,427 windows from 15 participants were available (182 stress and 1,245 non-stress); the corresponding per-participant window and class counts are reported in Supplementary Table S3. The annotations are used only for evaluation and never for model fitting.

For comparison with supervised baselines, we implemented logistic regression, random forest, and gradient boosting using standard R packages (`stats::glm` [38], `randomForest::randomForest` [39], and `gbm::gbm` [40]). The supervised classifiers were trained and evaluated under leave-one-subject-out (LOSO) cross-validation, whereas the proposed exceedance score is unsupervised and was evaluated in-sample (fitted once to each participant's full recording and scored on the same windows). Because the proposed score and the supervised classifiers are not evaluated under a common held-out design, the supervised AUCs are presented as *contextual benchmarks* that illustrate the performance of purely discriminative window-level methods,

rather than as a formal head-to-head comparison. The intended primary comparison for the proposed score is against the raw `HR_mean` baseline, which is evaluated under the same experimental design. These ensemble methods have shown strong discriminative performance in related translational classification tasks [41], but they do not model temporal dependence or subject-level baselines, in contrast to the proposed ARX–GARCHX risk-scoring model.

4.3 Results and subject-level insights

Using subject-specific thresholds $c_s(0.70)$ and Gaussian innovations, we evaluated the exceedance-based risk score $\widehat{\pi}_{s,t}$ against the raw 60-s heart-rate means (`HR_mean`). The pooled Receiver Operating Characteristic (ROC) analysis in Figure 4(A) shows that the variance-aware risk score provides substantially stronger separation across the annotated WESAD segments. The estimated AUC for the risk score was $\mathrm{AUC}_{\mathrm{ex}} = 0.931$, compared with $\mathrm{AUC}_{\mathrm{HR}} = 0.868$ for the raw heart-rate signal ($\Delta = 0.063$). Because repeated windows are clustered within only 15 participants, the pooled DeLong test may overstate precision. We therefore additionally report participant-clustered bootstrap confidence intervals obtained by resampling whole participants (3,000 bootstrap replicates). This yields 95% confidence intervals of [0.887, 0.972] for $\mathrm{AUC}_{\mathrm{ex}}$ and [0.821, 0.918] for $\mathrm{AUC}_{\mathrm{HR}}$, while the paired AUC difference has a 95% confidence interval of [0.031, 0.094], excluding zero. The improvement of the variance-aware risk score over the raw heart-rate signal is thus robust to within-participant clustering, indicating that modeling both mean and variance dynamics yields a more discriminative temporal representation than relying on raw heart-rate measurements alone.

For reference-based benchmarking, we also compared the proposed *unsupervised* metric with several *supervised* classifiers trained on window-level features using leave-one-subject-out cross-validation. Under LOSO cross-validation, logistic regression achieved $\mathrm{AUC}_{\mathrm{logit}} = 0.931$, random forest achieved $\mathrm{AUC}_{\mathrm{RF}} = 0.921$, and gradient boosting achieved $\mathrm{AUC}_{\mathrm{GBM}} = 0.948$. These supervised models use the annotations during training and are evaluated on held-out subjects, whereas the proposed score is unsupervised and evaluated in-sample; the two are therefore not evaluated under a common design, and we present the supervised values as contextual benchmarks rather than as a formal head-to-head comparison. Read in that light, the variance-aware score—estimated without any annotations—attains discrimination in the same range as these purely discriminative window-level classifiers, supporting its ability to extract informative temporal structure from multimodal wearable streams.

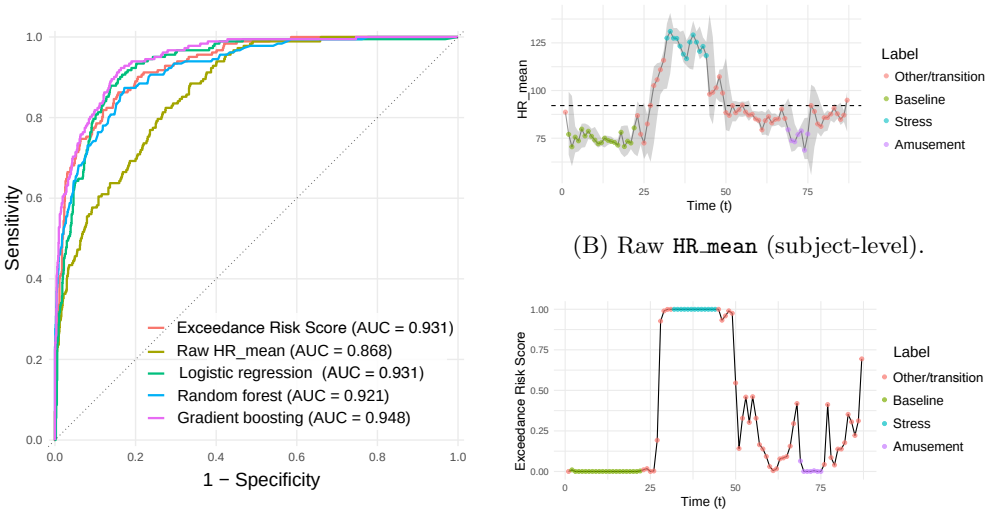
Two additional analyses examine the robustness of our results and identify the factors contributing to the observed performance. First, replacing the full-recording threshold $c_s(0.70)$ with a leak-free threshold estimated solely from an early burn-in window reduces $\mathrm{AUC}_{\mathrm{ex}}$ from 0.931 to approximately 0.880; notably, this remains at or above the raw `HR_mean` baseline (details in Supplementary Section S6). Second, we conduct a focused ablation to isolate the impact of covariate-driven variance by comparing Model B (no covariates in the variance equation) to Model D (the full model); this analysis, along with a calibration assessment for the exceedance event $\{y_{s,t} > c_s(0.70)\}$, is reported in Supplementary Section S7. Both variance-aware models substantially

outperform the raw `HR_mean` baseline and perform comparably to each other. This indicates that the real-data advantage is primarily driven by variance-aware modeling and is robust to the inclusion of covariates in the variance equation, whereas the benefits of covariate-driven variance are more pronounced in the structured-volatility simulation regimes.

It is worth noting that the supervised baselines treat all windows as independent samples and do not incorporate subject-specific baselines or temporal dependence. They therefore serve as strong discriminative benchmarks but do not capture the autoregressive or heteroscedastic structure exploited by the proposed temporal risk-scoring framework.

To complement the aggregate ROC analysis and highlight subject-level dynamics, we examine a representative individual trajectory. Figure 4(B) shows the raw 60-s `HR_mean` series for a representative participant (WESAD subject S11), with the subject-specific 70% threshold $c_s(0.70) \approx 92.10$ indicated by a dashed line. This participant was selected before qualitative inspection of the model output as a representative case whose protocol contains a clearly delineated stress segment; trajectories for all participants are provided in the replication materials. The annotated WESAD segments reveal a sustained elevation in heart rate during one region of the protocol, with lower variability elsewhere. Occasional excursions approach the threshold outside this segment but do not persist.

Figure 4(C) displays the corresponding risk score $\widehat{\pi}_{s,t}$. The score increases sharply at the onset of the high-variability segment, remains elevated throughout its duration, and drops promptly at its conclusion. Outside this interval, the score remains near zero despite moderate fluctuations in the raw signal. This behavior indicates that the variance-aware temporal model concentrates exceedance probability in periods characterized by sustained deviations relative to the subject's own baseline while suppressing short-lived fluctuations. Together, the two right-hand panels illustrate that the exceedance formulation provides a subject-adaptive, interpretable temporal summary that aligns closely with the annotated structure of the signal.



(A) ROC curves: pooled discrimination.

(B) Raw `HR_mean` (subject-level).

(C) Estimated risk score.

Fig. 4: WESAD results. **(A)** ROC curves comparing the variance-aware exceedance-based risk score with the raw 60-s heart-rate mean (`HR_mean`) and several supervised window-level baselines. The risk score and `HR_mean` are *unsupervised* (no annotation information used during model fitting) and evaluated in-sample, whereas logistic regression, random forest, and gradient boosting are *supervised* classifiers trained and evaluated using leave-one-subject-out cross-validation and shown as contextual benchmarks; the unsupervised risk score achieves AUC 0.931 (participant-clustered bootstrap 95% CI $[0.887, 0.972]$) and substantially improves upon the raw `HR_mean` baseline (0.868; difference 95% CI $[0.031, 0.094]$). **(B)** Subject-level raw 60-s `HR_mean` for one representative participant with the subject-specific 70% threshold $c_s(0.70) \approx 92.10$ shown as a dashed line. **(C)** The corresponding estimated risk score over time, showing elevated values during the annotated stress-associated segment and near-zero values elsewhere.

The risk score offers a principled and interpretable way to summarize temporal deviations in wearable sensor streams. Its probabilistic range between 0 and 1 provides a normalized scale that is comparable across subjects and robust to individual baseline differences. Unlike raw heart-rate thresholds, the score is sensitive to sustained departures yet resistant to transient fluctuations, yielding a stable temporal characterization of signal patterns. Importantly, the score is estimated without using any annotations, yet it naturally differentiates periods of elevated and quiescent activity in the WESAD protocol. This highlights its utility as an unsupervised temporal feature, suitable both for standalone interpretation and as a noise-robust, low-dimensional input to downstream sensing or machine-learning models.

4.4 Computational feasibility

The WESAD analysis involved fifteen participants, approximately one hundred twenty windows per participant, and seventy-seven covariates. With twenty outer iterations, the mean-update block ran in roughly 0.01 s per iteration and the variance-update block in roughly 0.12 s per iteration, so end-to-end training completed in approximately 26.6 s. After training, a new-window score was computed with sub-millisecond latency. Detailed timings, asymptotic complexity ($O(STd)$ for the mean recursion and $O(ST(L+R+M))$ for the variance recursion), memory footprint, and hardware information are reported in Supplementary Table S2 (Supplementary Section S4). These results support potential use as a lightweight feature-extraction or monitoring component, although clinical deployment would require prospective validation.

5 Discussion

This study reframes variance-aware wearable-sensor modeling as a translational temporal risk-detection problem. The main finding is that explicitly modeling conditional variance, rather than treating variability as nuisance noise, yields an interpretable subject-adaptive score that improves discrimination of stress-associated physiological windows relative to raw heart-rate summaries. In WESAD, the risk score achieved an AUC of 0.931 without using labels during model fitting—compared with 0.868 for the raw heart-rate mean, a difference that remains significant under a participant-clustered bootstrap—while attaining discrimination in the same range as supervised logistic-regression, random-forest, and gradient-boosting classifiers evaluated as contextual benchmarks, and remaining computationally lightweight. We caution that the supervised classifiers were trained and evaluated under leave-one-subject-out cross-validation whereas the unsupervised score was evaluated in-sample, so this correspondence is contextual rather than a formal head-to-head comparison.

The method has several translationally relevant properties. First, it produces a probability-scale score between 0 and 1, which is easier to interpret than arbitrary transformed features. Second, it uses individualized thresholds, making the score more robust to baseline differences across participants. Third, it incorporates multimodal covariates into both the mean and variance equations, allowing the model to distinguish sustained physiological shifts from transient fluctuations. Fourth, it can be computed rapidly after training, which is important for real-time or near-real-time wearable monitoring.

The approach also complements supervised machine learning. Deep learning and sequence models may achieve strong classification accuracy when large labeled datasets are available [42, 43], but many translational applications lack dense, reliable labels. The proposed score can be used as an unsupervised or weakly supervised feature that summarizes subject-specific risk dynamics before downstream prediction. It may therefore be useful as a digital biomarker component in studies of stress, autonomic dysregulation, symptom exacerbation, or other temporally dynamic states. The same idea may also connect with broader time-series segmentation and change-detection methods for multivariate data [44].

Several limitations should be emphasized. The WESAD analysis is a controlled laboratory benchmark with a small number of participants and should not be interpreted as clinical validation. Stress labels in WESAD represent protocol-defined affective states rather than clinically diagnosed stress disorders. External validation in larger and more diverse cohorts is needed before any medical claim can be made. In addition, the supervised classifiers were evaluated under leave-one-subject-out cross-validation, while the risk score was fitted without labels and evaluated post hoc; a fully prospective validation should freeze preprocessing, tuning, thresholds, and model parameters within an outer held-out-subject or external-cohort protocol. The current model also uses fixed lag orders and a balanced-panel formulation. Irregular sampling, missingness, longer-range temporal dependence, and nonlinear mean-variance effects remain important extensions.

Finally, the risk score is threshold-dependent. While we used the 70th percentile of each participant's heart-rate distribution for interpretability and cross-participant comparison, alternative thresholds may be better suited for different clinical or behavioral endpoints. Because this quantile is computed over the entire recording, it inherently incorporates the same elevated heart-rate windows that later define the positive class. Although no labels are used, this may introduce a modest optimistic bias to the reported discrimination. A sensitivity analysis fixing c_s from a leak-free early burn-in window (Supplementary Section S6) reduces $\mathrm{AUC}_{\mathrm{ex}}$ from 0.931 to approximately 0.880 but remains at or above the raw heart-rate baseline. This confirms that the temporal signal is genuine while emphasizing that the exact magnitude of performance is sensitive to threshold selection. In future translational applications, thresholds should be fixed prospectively from an initial calibration period and chosen from clinical anchors, patient-reported outcomes, or adverse-event definitions, rather than relying solely on empirical quantiles computed over the full recording.

6 Conclusions

A penalized panel ARX–GARCHX model can convert high-dimensional wearable physiological streams into an interpretable, subject-adaptive temporal risk score. In simulations, variance-aware modeling improved recovery of threshold-exceedance risk under structured volatility. In the WESAD benchmark, the score improved stress-associated window discrimination relative to raw heart rate and performed competitively with supervised baselines while requiring less than 1 ms for new-window inference after training. The method is a promising digital biomarker workflow for translational wearable studies, but prospective external validation is required before clinical use.

List of abbreviations

ACC	acceleration
ARX	autoregressive model with exogenous covariates
AUC	area under the receiver operating characteristic curve
BRV	breathing-rate variability
ECG	electrocardiogram
EDA	electrodermal activity
EMG	electromyography
GARCHX	generalized autoregressive conditional heteroscedasticity model with exogenous covariates
HR	heart rate
HRV	heart-rate variability
QMLE	quasi-maximum-likelihood estimation
RESP	respiration
RMSE	root mean squared error
ROC	receiver operating characteristic
WESAD	Wearable Stress and Affect Detection

Supplementary information

The Supplementary Material is organized into eight sections. Supplementary Section S1 (*Theoretical Results*) states the regularity, restricted-strong-convexity, and score-bound assumptions and proves a finite-sample oracle inequality (Theorem S1) showing that the penalized estimator achieves an excess-risk rate of order $(s_\beta + s_\gamma) \log p/(ST)$. Supplementary Section S2 (*Algorithmic Details*) gives the full estimation pseudocode (Algorithm S1) for the penalized panel ARX–GARCHX model, including the weighted closed-form mean update, the decoupled (ω, a, b) and γ variance updates, positivity safeguards, and tuning conventions. Supplementary Section S3 (*Supplementary Simulation Results*) reports the replication-level Bias and RMSE summaries for Models A–F under Scenarios S1–S4 at $T \in \{120, 240\}$ in Supplementary Table S1. Supplementary Section S4 (*Supplementary Computational Results*) reports the runtime profile, asymptotic complexity, memory footprint, and hardware specification for the WESAD analysis in Supplementary Table S2. Supplementary Section S5 (*WESAD Endpoint and Per-Participant Composition*) defines the evaluation endpoint, tabulates per-participant window and class counts (Table S3), and reports participant-clustered bootstrap confidence intervals. Supplementary Section S6 (*Sensitivity to the Threshold Choice*) reports the leak-free threshold analysis. Supplementary Section S7 (*WESAD Variance-Covariate Ablation*) compares Model B and Model D on discrimination and calibration of the exceedance event. Supplementary Section S8 (*Reproducibility Details*) lists the complete 77-feature set (Table S4), retained windows, quality control, tuning grid, temporal folds, and software/code provenance.

Declarations

Ethics approval and consent to participate

The present study was a secondary analysis of the publicly available, de-identified WESAD dataset and did not involve new recruitment, intervention, or data collection from human participants. Ethical and consent procedures for the original WESAD data collection were reported by the original dataset investigators [4]. No additional ethics approval was required for this secondary analysis.

Consent for publication

Not applicable.

Trial registration

Not applicable. This study is a methodological article that uses publicly available secondary data and does not constitute a clinical trial.

Reporting guidelines

Not applicable in the formal sense. This is a methodological article using publicly available secondary data; established clinical reporting checklists such as TRIPOD (for prediction-model development and validation) and STARD (for diagnostic accuracy) do not directly apply because the analysis was a proof-of-concept feature-extraction demonstration on a public benchmark, with no prospective enrollment, diagnostic claim, or external clinical validation. Where relevant, we have followed the spirit of the TRIPOD-AI items by reporting the data source, preprocessing pipeline, model specification, tuning procedure, evaluation metrics, intended use, and limitations.

Availability of data and materials

The WESAD dataset analyzed in this study is publicly available through the UC Irvine Machine Learning Repository and is described in the original dataset publication [4, 33].

Code availability

An R package implementing the proposed framework, `varGuidTS`, which builds on the variance-guided regression framework of [32], is available at https://cloud.r-project.org/web/packages/varGuidTS/. Additional source code and replication materials for the analyses reported in this paper are available at https://github.com/zionwzz/variance-guided-risk-demo.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35 GM139659, the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01 HL164405, and the 2024 Relief Funding Award from the Office of the Vice Provost for Research and Scholarship and the Office of Faculty Affairs, University of Miami. The funders had no role in study design, data analysis, interpretation of results, or preparation of the manuscript.

Authors' contributions

ZW contributed to methodology, software implementation, data preprocessing, simulation studies, real-data analysis, visualization, and manuscript drafting. ML contributed to conceptualization, methodology, supervision, funding acquisition, interpretation of results, and manuscript revision. Both authors read and approved the final manuscript.

Acknowledgements

The authors thank the WESAD investigators for making the dataset publicly available.

Use of AI-assisted tools

During preparation of this journal-specific draft, the authors used an AI language model to assist with editorial restructuring and wording. The authors reviewed and are responsible for all scientific content, analyses, interpretations, and conclusions.

References

[1] Smets E, Rios Velazquez E, Schiavone G, Chakroun I, D'Hondt E, De Raedt W, et al. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. NPJ digital medicine. 2018;1(1):67.

[2] Valenza G, Nardelli M, Lanata A, Gentili C, Bertschy G, Paradiso R, et al. Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis. IEEE Journal of Biomedical and Health Informatics. 2013;18(5):1625–1635.

[3] Stan IE, D'Auria D, Napoletano P. A Systematic Literature Review of Innovations, Challenges, and Future Directions in Telemonitoring and Wearable Health Technologies. IEEE Journal of Biomedical and Health Informatics. 2025;.

[4] Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction; 2018. p. 400–408.

[5] Assabumrungrat R, Sangnark S, Charoenpattarawut T, Polpakdee W, Sudhawiyangkul T, Boonchieng E, et al. Ubiquitous affective computing: A review. IEEE Sensors Journal. 2021;22(3):1867–1881.

[6] Abdelfattah E, Joshi S, Tiwari S. Machine and deep learning models for stress detection using multimodal physiological data. IEEE Access. 2025;.

[7] Liu Y, Zavareh AT, Zoghi B. Enhancing Well-Being and Alleviating Stress via Wearable-Driven Emotion Recognition and EQ Intentional Practice with Deep Reinforcement Learning. IEEE Sensors Letters. 2024;.

[8] Tian F, Zhang L, Zhu L, Zhao M, Liu J, Dong Q, et al. Advancements in affective disorder detection: Using multimodal physiological signals and neuromorphic computing based on snns. IEEE Transactions on Computational Social Systems. 2024;.

[9] Vidyasagar KC, Kumar KR, Sai GA, Ruchita M, Saikia MJ. Signal to image conversion and convolutional neural networks for physiological signal processing: A review. Ieee Access. 2024;12:66726–66764.

[10] Minor B, Greeley C, Holder R, Thomas B, Holder LB, Cook DJ. A Feature-Augmented Transformer Model to Recognize Functional Activities from in-the-wild Smartwatch Data. IEEE Journal of Biomedical and Health Informatics. 2025;.

[11] Arif S, Siddiqui MM, Alqahtani N, Shah K, Khan MA, Kraiem N. Advanced Signal Processing Algorithm-Based State Estimation and Denoising of ECG Signals for Biomedical Applications. IEEE Access. 2025;.

[12] Lee S, Lee S, Jang W, Kim J, Yon DK, Lee J. Weighted Iterative Complex Demodulation for High-Resolution Instantaneous Frequencies in Low-Frequency PPG Signals from Wearable Devices. IEEE Journal of Biomedical and Health Informatics. 2025;.

[13] Tian Y, Wei H, Tan J. An adaptive-gain complementary filter for real-time human motion tracking with MARG sensors in free-living environments. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2012;21(2):254–264.

[14] Ardalan S, Moghadami S, Jaafari S. Motion noise cancelation in heartbeat sensing using accelerometer and adaptive filter. IEEE Embedded Systems Letters. 2015;7(4):101–104.

[15] De Sabbata G, Simonini G. Real-Time Forecasting from Wearable-Monitored Heart Rate Data Through Autoregressive Models. Journal of Healthcare Informatics Research. 2025;p. 1–20.

[16] Tavares T, Nogueira M, Rosário D, Santos A, Cerqueira E. Traffic model based on autoregression for ppg signals in wearable networks. IEEE Networking Letters. 2020;2(2):49–53.

[17] Salem O, Alsubhi K, Mehaoua A, Boutaba R. Markov models for anomaly detection in wireless body area networks for secure health monitoring. IEEE Journal on Selected Areas in Communications. 2020;39(2):526–540.

[18] Fathian R, Khandan A, Rahmanifar N, Ho C, Rouhani H. Feasibility and validity of wearable sensors for monitoring temporal parameters in manual wheelchair propulsion. IEEE Journal of Biomedical and Health Informatics. 2024;28(9):5239–5246.

[19] Ben-Moshe N, Tsutsui K, Brimer SB, Zvuloni E, Sörnmo L, Behar JA. RawECGNet: Deep learning generalization for atrial fibrillation detection from the raw ECG. IEEE Journal of Biomedical and Health Informatics. 2024;28(9):5180–5188.

[20] Leng J, Li H, Shi W, Gao L, Lv C, Wang C, et al. Time-frequency-space EEG decoding model based on dense graph convolutional network for stroke. IEEE Journal of Biomedical and Health Informatics. 2024;28(9):5214–5226.

[21] Ljung L. System Identification: Theory for the User. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 1999.

[22] Engle RF. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica. 1982;50(4):987–1007.

[23] Han H, Kristensen D. Asymptotic theory for the QMLE in GARCH-X models with stationary and nonstationary covariates. Journal of business & economic statistics. 2014;32(3):416–429.

[24] Tse YK. A test for constant correlations in a multivariate GARCH model. Journal of econometrics. 2000;98(1):107–127.

[25] Yu C, Li D, Jiang F, Zhu K. Matrix GARCH model: Inference and application. Journal of the American Statistical Association. 2025;120(551):1747–1762.

[26] Zhu Q, Tan S, Zheng Y, Li G. Quantile autoregressive conditional heteroscedasticity. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2023;85(4):1099–1127.

[27] Blundell R, Bond S. Initial conditions and moment restrictions in dynamic panel data models. Journal of econometrics. 1998;87(1):115–143.

[28] Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267–288.

[29] Zou H. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association. 2006;101(476):1418–1429.

[30] Wang H, Li G, Tsai CL. Regression coefficient and autoregressive order shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2007;69(1):63–78.

[31] Medeiros MC, Mendes EF. L1-Regularization of High-Dimensional Time-Series Models with Non-Gaussian and Heteroskedastic Innovations. Journal of Econometrics. 2016;191(1):150–164.

[32] Liu S, Lu M. Variance-Guided Regression for Heteroscedastic Data with a Grouping-Based Extension for Nonlinear Prediction. Statistics in Medicine. 2026;45(13-14):e70632. https://doi.org/10.1002/sim.70632.

[33] UC Irvine Machine Learning Repository.: WESAD (Wearable Stress and Affect Detection). Accessed 14 May 2026. Available from: https://archive.ics.uci.edu/ml/datasets/WESAD+%28Wearable+Stress+and+Affect+Detection%29.

[34] MacNab Y, Dean C. Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models. Statistics in medicine. 2000;19(17-18):2421–2435.

[35] Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association. 2001;96(456):1348–1360. https://doi.org/10.1198/016214501753382273.

[36] Zhang CH. Nearly Unbiased Variable Selection under Minimax Concave Penalty. The Annals of Statistics. 2010;38(2):894–942. https://doi.org/10.1214/09-AOS729.

[37] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior research methods. 2021;53(4):1689–1696.

[38] R Core Team.: R: A Language and Environment for Statistical Computing. Vienna, Austria. Available from: https://www.R-project.org/.

[39] Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

[40] Ridgeway G, GBM Developers.: gbm: Generalized Boosted Regression Models. R package version 2.2.2. Available from: https://CRAN.R-project.org/package=gbm.

[41] Lu M, Yin R, Chen XS. Ensemble Methods of Rank-Based Trees for Single Sample Classification with Gene Expression Profiles. Journal of Translational Medicine.

2024;22(140). https://doi.org/10.1186/s12967-024-04940-2.

[42] Wang Z, Wang Y, Hu C, Yin Z, Song Y. Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model. IEEE Sensors Journal. 2022;22(5):4359–4368.

[43] Han Z, Zhao J, Leung H, Ma KF, Wang W. A review of deep learning models for time series prediction. IEEE Sensors Journal. 2019;21(6):7833–7848.

[44] McGonigle ET, Cho H. Nonparametric data segmentation in multivariate time series via joint characteristic functions. Biometrika. 2025;112(2):asaf024.

SUPPLEMENTARY MATERIAL

Variance-Aware Penalized Panel Models for Temporal Risk Detection from Wearable Sensor Data

S1 Theoretical Results

Under standard smoothness, dependence, and identifiability conditions (Assumptions S1–S3 below), the penalized ARX–GARCHX estimator $(\widehat{\beta}, \widehat{\gamma})$ attains estimation error, up to a constant factor, comparable to that of the best sparse oracle choice of (β, γ). In particular, the estimator correctly identifies the leading covariates in both the mean and variance equations and yields stable mean–variance estimates in high-dimensional multisubject settings. The rest of this section formalizes the assumptions, states the oracle inequality (Theorem S1), and gives a self-contained proof.

Assumption S1 (Regularity and dependence)**.** *The panel process* $\{(y_{s,t}, x_{s,t})\}$ *satisfies standard mixing and moment conditions, and the loss* $\ell_{s,t}(\beta, \gamma)$ *is Lipschitz and twice differentiable in* (β, γ) *with bounded derivatives. These conditions ensure that* $R(\beta, \gamma) = \mathbb{E}[\ell_{s,t}(\beta, \gamma)]$ *is well defined and smooth, and that empirical-score deviations concentrate at rate* $\sqrt{(\log p)/(ST)}$*, where* p *denotes the ambient dimension of the penalized parameters. In our setting,* $x_{s,t} \in \mathbb{R}^d$*, and the penalized mean and variance coefficients* (β, γ) *each lie in* $\mathbb{R}^d$*, so* p *is of order* d*.*

Assumption S2 (Restricted strong convexity (RSC))**.** *There exist constants* $\kappa > 0$ *and* $\tau \geq 0$ *such that, for all* $(\Delta_\beta, \Delta_\gamma)$ *in the usual cone associated with the oracle supports* $S_\beta = \mathrm{supp}(\beta^\star)$ *and* $S_\gamma = \mathrm{supp}(\gamma^\star)$*,*

$$\left\langle \nabla R(\beta^\star + \Delta_\beta, \gamma^\star + \Delta_\gamma) - \nabla R(\beta^\star, \gamma^\star), (\Delta_\beta, \Delta_\gamma) \right\rangle \geq \kappa(\|\Delta_\beta\|_2^2 + \|\Delta_\gamma\|_2^2) - \tau \frac{\log p}{ST}.$$

Assumption S3 (Score bound)**.** *The empirical score at the oracle pair satisfies*

$$\left\| \frac{1}{ST} \sum_{s,t} \nabla_{\beta,\gamma} \ell_{s,t}(\beta^\star, \gamma^\star) \right\|_\infty \leq c_0 \sqrt{\frac{\log p}{ST}}$$

with probability at least $1 - Ce^{-c \log p}$ *for constants* $c, C > 0$.

Theorem S1 (Oracle inequality for penalized ARX–GARCHX)**.** *Let* $(\widehat{\beta}, \widehat{\gamma})$ *minimize the penalized objective*

$$(\widehat{\beta}, \widehat{\gamma}) \in \arg\min_{\beta,\gamma} \Big\{ \mathcal{Q}_{S,T}(\beta, \gamma) := \frac{1}{ST} \sum_{s=1}^{S} \sum_{t=1}^{T} \ell_{s,t}(\beta, \gamma) + \lambda_\beta \|\beta\|_1 + \lambda_\gamma \|\gamma\|_1 \Big\}, \tag{1}$$

and let $(\beta^\star, \gamma^\star)$ *be the oracle sparse pair with* $\|\beta^\star\|_0 \leq s_\beta$ *and* $\|\gamma^\star\|_0 \leq s_\gamma$*. Under Assumptions S1–S3, and for tuning parameters* $\lambda_\beta, \lambda_\gamma \asymp \sqrt{(\log p)/(ST)}$*, we have, with probability at least* $1 - C \exp(-c \log p)$*,*

$$R(\widehat{\beta}, \widehat{\gamma}) - R(\beta^\star, \gamma^\star) \lesssim (s_\beta + s_\gamma) \frac{\log p}{ST}, \tag{2}$$

where the hidden constant depends only on (κ, τ, c_0)*.*

Proof. The proof follows a standard high-dimensional M-estimation argument, adapted to the panel time-series setting via the assumptions on dependence and RSC.

Step 1: Basic inequality. By definition of $(\widehat{\beta}, \widehat{\gamma})$ in (1) and feasibility of $(\beta^\star, \gamma^\star)$, we have

$$\begin{aligned} & \mathcal{Q}_{S,T}(\widehat{\beta}, \widehat{\gamma}) \leq \mathcal{Q}_{S,T}(\beta^\star, \gamma^\star) \\ \Rightarrow \quad & \frac{1}{ST} \sum_{s,t} \big\{ \ell_{s,t}(\widehat{\beta}, \widehat{\gamma}) - \ell_{s,t}(\beta^\star, \gamma^\star) \big\} + \lambda_\beta \big(\|\widehat{\beta}\|_1 - \|\beta^\star\|_1 \big) + \lambda_\gamma \big(\|\widehat{\gamma}\|_1 - \|\gamma^\star\|_1 \big) \leq 0. \end{aligned}$$

Define $\Delta_\beta = \widehat{\beta} - \beta^\star$ and $\Delta_\gamma = \widehat{\gamma} - \gamma^\star$, and write the empirical risk difference as

$$\underbrace{\{R(\widehat{\beta}, \widehat{\gamma}) - R(\beta^\star, \gamma^\star)\}}_{\text{population}} + \underbrace{\Big[\tfrac{1}{ST} \sum_{s,t} \big(\ell_{s,t}(\widehat{\beta}, \widehat{\gamma}) - \ell_{s,t}(\beta^\star, \gamma^\star) \big) - \big(R(\widehat{\beta}, \widehat{\gamma}) - R(\beta^\star, \gamma^\star) \big) \Big]}_{\text{empirical process}} \leq -\lambda_\beta \big(\|\widehat{\beta}\|_1 - \|\beta^\star\|_1 \big) - \lambda_\gamma \big(\|\widehat{\gamma}\|_1 - \|\gamma^\star\|_1 \big).$$

Step 2: Controlling the empirical process term. By Assumptions S1 and S3, which provide smoothness and score concentration under weak dependence, the empirical process term

satisfies, on an event of probability at least $1 - C\exp(-c\log p)$,

$$\Big|\tfrac{1}{ST}\sum_{s,t}\big(\ell_{s,t}(\widehat{\beta},\widehat{\gamma}) - \ell_{s,t}(\beta^\star,\gamma^\star)\big) - \big(R(\widehat{\beta},\widehat{\gamma}) - R(\beta^\star,\gamma^\star)\big)\Big| \;\le\; \tfrac{\lambda_\beta}{2}\|\Delta_\beta\|_1 + \tfrac{\lambda_\gamma}{2}\|\Delta_\gamma\|_1,$$

for $\lambda_\beta, \lambda_\gamma \asymp \sqrt{(\log p)/(ST)}$ up to constants depending on c_0. Substituting this bound into the basic inequality yields

$$\begin{aligned} R(\widehat{\beta},\widehat{\gamma}) - R(\beta^\star,\gamma^\star) &\le \tfrac{\lambda_\beta}{2}\|\Delta_\beta\|_1 + \tfrac{\lambda_\gamma}{2}\|\Delta_\gamma\|_1 - \lambda_\beta\big(\|\widehat{\beta}\|_1 - \|\beta^\star\|_1\big) - \lambda_\gamma\big(\|\widehat{\gamma}\|_1 - \|\gamma^\star\|_1\big) \\ &\le \tfrac{3\lambda_\beta}{2}\|\Delta_{\beta,S_\beta}\|_1 + \tfrac{3\lambda_\gamma}{2}\|\Delta_{\gamma,S_\gamma}\|_1, \end{aligned}$$

where in the last inequality we used the usual cone constraints $\|\Delta_{\beta,S_\beta^c}\|_1 \le 3\|\Delta_{\beta,S_\beta}\|_1$ and similarly for γ, which follow from standard ℓ_1-penalized estimation arguments.

Step 3: Restricted strong convexity and oracle rate. By Assumption S2 (RSC), we have

$$R(\widehat{\beta},\widehat{\gamma}) - R(\beta^\star,\gamma^\star) \;\ge\; \kappa\big(\|\Delta_\beta\|_2^2 + \|\Delta_\gamma\|_2^2\big) - \tau\frac{\log p}{ST}.$$

Combining this with the upper bound from Step 2 and using the inequalities $\|\Delta_{\beta,S_\beta}\|_1 \le \sqrt{s_\beta}\|\Delta_\beta\|_2$ and $\|\Delta_{\gamma,S_\gamma}\|_1 \le \sqrt{s_\gamma}\|\Delta_\gamma\|_2$, we obtain

$$\kappa\big(\|\Delta_\beta\|_2^2 + \|\Delta_\gamma\|_2^2\big) \;\lesssim\; \lambda_\beta\sqrt{s_\beta}\|\Delta_\beta\|_2 + \lambda_\gamma\sqrt{s_\gamma}\|\Delta_\gamma\|_2 + \tau\frac{\log p}{ST}.$$

A standard quadratic inequality then yields

$$\|\Delta_\beta\|_2^2 + \|\Delta_\gamma\|_2^2 \;\lesssim\; (s_\beta + s_\gamma)\,\lambda^2 \;\asymp\; (s_\beta + s_\gamma)\,\frac{\log p}{ST},$$

where λ^2 stands for the common order of λ_β^2 and λ_γ^2. Substituting back into the RSC bound on the excess risk gives

$$R(\widehat{\beta},\widehat{\gamma}) - R(\beta^\star,\gamma^\star) \;\lesssim\; (s_\beta + s_\gamma)\,\frac{\log p}{ST},$$

which is exactly (2). This completes the proof. □

S2 Algorithmic Details

The main manuscript introduces the penalized panel ARX–GARCHX model and the exceedance-based risk score. This section gives the full estimation algorithm and the implementation details used for the simulation and WESAD analyses.

The estimator alternates between a mean block and a variance block via block coordinate descent. The mean block updates the subject-specific intercepts α_s in closed form and then the shared mean parameters (θ, β) via a weighted penalized regression step. The variance block is decoupled in two stages: a constrained nonlinear update for the GARCH parameters (ω, a, b) with γ fixed (B1), followed by a proximal-gradient update for the variance covariate effects γ with (ω, a, b) fixed (B2). A forward recursion then refreshes the conditional variance $\sigma^2_{s,t}$ for the next outer iteration. The full procedure is summarized in Algorithm 1.

Subject-specific intercepts. Given current conditional variance estimates, the subject-specific intercepts in the conditional mean equation are updated by weighted least squares:

$$\alpha_s \leftarrow \frac{\sum_t w_{s,t}\left(y_{s,t} - \sum_{i=1}^{P} \theta_i y_{s,t-i} - \sum_{j=0}^{Q} x_{s,t-j}^{\top}\beta_j\right)}{\sum_t w_{s,t}}, \qquad w_{s,t} = \sigma_{s,t}^{-2}. \tag{3}$$

A centering constraint $\sum_s \alpha_s = 0$ may be imposed for identifiability.

Convexity and positivity safeguards. Each block is convex once the others are held fixed. The mean block updates the subject-specific intercepts α_s in closed form and then the shared mean parameters (θ, β) via a weighted penalized regression step. The variance update is convex in the non-negative GARCH parameters (ω, a, b) for fixed γ (subblock B1) and convex in γ for fixed (ω, a, b) (subblock B2), since $u \mapsto \log u + c/u$ is convex on $(0, \infty)$. Positivity of the conditional variance is enforced by constraining $\omega_s > 0$, $a_r \geq 0$, and $b_\ell \geq 0$, and by projecting the variance recursion onto a small positive lower bound inside the log and ratio terms whenever the unconstrained linear predictor would otherwise approach zero.

Implementation choices used in this paper. Covariates were standardized within subject before model fitting. The primary analyses used lag orders $(P, Q, L, R, M) = (1, 0, 1, 1, 0)$, reflecting the short-memory structure of typical wearable sensor data; sensitivity checks with larger orders did not materially change the results. The penalties (λ, ρ) for the mean and variance blocks were selected by blockwise time-series cross-validation, partitioning each subject's

Algorithm 1 Penalized Panel ARX–GARCHX with Decoupled Variance Updates

Input: Panel $\{(y_{s,t}, x_{s,t})\}$; orders P, Q (mean), L, R, M (variance); penalties $P_\lambda(\beta), P_\rho(\gamma)$.

Init: $\Theta^{(0)} = (\alpha^{(0)}, \theta^{(0)}, \beta^{(0)})$, $\Phi^{(0)} = (\omega^{(0)}, a^{(0)}, b^{(0)}, \gamma^{(0)})$; compute $\sigma^{2,(0)}_{s,t}$.

for $k = 0, 1, \ldots$ until convergence **do**

(A) Mean block (weighted penalized regression).

$w_{s,t} \leftarrow 1/\sigma^{2,(k)}_{s,t}$.

$$\alpha_s^{(k+1)} \leftarrow \frac{\sum_t w_{s,t}\big(y_{s,t} - \sum_i \theta_i^{(k)} y_{s,t-i} - \sum_j \beta_j^{(k)} x_{s,t-j}\big)}{\sum_t w_{s,t}}.$$

$$\{\theta, \beta\}^{(k+1)} \in \arg\min_{\theta,\beta} \sum_{s,t} w_{s,t}\big(y_{s,t} - \alpha_s^{(k+1)} - \sum_i \theta_i y_{s,t-i} - \sum_j \beta_j x_{s,t-j}\big)^2 + P_\lambda(\beta).$$

$\hat{e}^{(k+1)}_{s,t} \leftarrow y_{s,t} - \mu_{s,t}(\alpha^{(k+1)}, \theta^{(k+1)}, \beta^{(k+1)})$.

Build strict mask

$I = \{(s,t) : \text{all required lags of } y, e, \sigma^2, x \text{ exist}\}$.

(B1) Variance sub-block: (ω, a, b) **given** $\gamma^{(k)}$**.**

$\delta_{s,t}(\omega, a, b) := \omega_s + \sum_{r=1}^R a_r\, \hat{e}^{2,(k+1)}_{s,t-r} + \sum_{\ell=1}^L b_\ell\, \sigma^{2,(k)}_{s,t-\ell}$.

$$(\omega^\star, a^\star, b^\star) \in \arg\min_{\omega,\, a,b\geq 0,\, \sum a + \sum b \leq \phi_{\max}} \sum_{(s,t)\in I} \Big\{ \log u_{s,t} + \hat{e}^{2,(k+1)}_{s,t}/u_{s,t} \Big\},$$

where $u_{s,t} = \delta_{s,t}(\omega, a, b) + x_{s,t}^\top \gamma^{(k)}$.

(B2) Variance sub-block: γ **given** $\delta^\star$**.**

With $\delta^\star_{s,t} = \delta_{s,t}(\omega^\star, a^\star, b^\star)$ fixed, $\gamma^{(k+1)} \in \arg\min_\gamma \sum_{(s,t)\in I} \Big\{ \log(\delta^\star_{s,t} + x_{s,t}^\top\gamma) + \hat{e}^{2,(k+1)}_{s,t}/(\delta^\star_{s,t} + x_{s,t}^\top\gamma) \Big\} + P_\rho(\gamma)$.

(C) Forward recursion: $\sigma^{2,(k+1)}_{s,t} \leftarrow \delta^\star_{s,t} + x_{s,t}^\top \gamma^{(k+1)}$.

end for

Output: $\widehat{\alpha}_s, \widehat{\theta}, \widehat{\beta}, \widehat{\omega}_s, \widehat{a}, \widehat{b}, \widehat{\gamma}$.

windows into contiguous folds to preserve temporal ordering and aggregating validation losses across subjects. Convergence was declared when the relative change in the full objective fell below a prespecified tolerance or after the maximum number of outer iterations (20 in the WESAD analysis), at which point fewer than 15 outer iterations were typically sufficient.

S3 Supplementary Simulation Results

Table S1 reports the replication-level Bias and RMSE of the estimated exceedance risk for all six competing models. Columns A–D are the original variants; **Model E** (ARX_p–GARCHX; penalized mean, unpenalized X-dependent variance) and **Model F** (AR–GARCH; no covariates in the mean or variance) were added in response to review. We compute E and F under the identical generator, evaluator, tuning, and seed scheme as the reported A–D run (100 replications), so that Columns A–D are unchanged and the two new columns are directly comparable.

S4 Supplementary Computational Results

Table S2 reports the runtime profile for the WESAD analysis. These details are provided in the supplement so that the main manuscript remains focused on the methodology and translational demonstration.

Table S2: Computational characteristics of the temporal risk-score model in the WESAD analysis.

Component	**Metric**	**Value**
Dataset scale	(S, T, d)	$(15, 120, 77)$
Mean update	Time per iteration	≈ 0.01 s
Variance update	Time per iteration	≈ 0.12 s
Total training	End-to-end runtime	≈ 26.6 s for 20 iterations
Inference	Latency for one new window	< 1 ms
Asymptotic complexity	Mean and variance recursions	$O(STd)$ and $O(ST(L + R + M))$
Memory footprint	Space complexity	$O(ST + d)$

Timings were obtained on a MacBook Air with an M2 processor and eight GB RAM. These timings document practical feasibility for the WESAD benchmark rather than optimized software performance.

Table S1: Bias and RMSE of estimated exceedance risk (means and standard errors in parentheses) across simulation Scenarios 1–4 with 100 replications, for the six competing models.

(a) Bias

T=120

	Model					
Scenario	A ARX–GARCH	B ARXp–GARCH	C ARX–GARCHX	D ARXp–GARCHXp	E ARXp–GARCHX	F AR–GARCH
1	-0.001 (0.001)	-0.002 (0.001)	0.004 (0.001)	0.005 (0.002)	-0.003 (0.001)	-0.001 (0.001)
2	-0.002 (0.000)	-0.002 (0.001)	-0.001 (0.000)	-0.001 (0.001)	-0.002 (0.001)	-0.010 (0.002)
3	-0.000 (0.002)	-0.000 (0.002)	0.005 (0.002)	0.004 (0.001)	-0.002 (0.001)	-0.002 (0.001)
4	-0.005 (0.000)	-0.005 (0.000)	0.001 (0.001)	-0.000 (0.001)	-0.004 (0.000)	-0.015 (0.002)

T=240

	Model					
Scenario	A ARX–GARCH	B ARXp–GARCH	C ARX–GARCHX	D ARXp–GARCHXp	E ARXp–GARCHX	F AR–GARCH
1	-0.001 (0.001)	-0.001 (0.001)	0.014 (0.001)	0.014 (0.001)	-0.002 (0.000)	-0.003 (0.002)
2	-0.000 (0.000)	-0.001 (0.000)	0.001 (0.000)	0.000 (0.001)	-0.002 (0.000)	-0.006 (0.001)
3	-0.002 (0.000)	-0.003 (0.001)	0.012 (0.001)	0.012 (0.001)	-0.004 (0.000)	-0.002 (0.002)
4	-0.005 (0.000)	-0.005 (0.000)	0.001 (0.000)	0.000 (0.001)	-0.005 (0.000)	-0.015 (0.001)

(b) RMSE

T=120

	Model					
Scenario	A ARX–GARCH	B ARXp–GARCH	C ARX–GARCHX	D ARXp–GARCHXp	E ARXp–GARCHX	F AR–GARCH
1	0.085 (0.001)	0.084 (0.001)	**0.080 (0.001)**	0.082 (0.003)	0.082 (0.002)	0.161 (0.002)
2	0.072 (0.001)	0.073 (0.002)	0.064 (0.001)	**0.063 (0.002)**	0.068 (0.002)	0.107 (0.003)
3	0.096 (0.003)	0.093 (0.003)	0.087 (0.003)	**0.084 (0.002)**	0.090 (0.003)	0.133 (0.001)
4	0.093 (0.000)	0.091 (0.000)	0.094 (0.001)	0.093 (0.002)	**0.089 (0.001)**	0.127 (0.002)

T=240

	Model					
Scenario	A ARX–GARCH	B ARXp–GARCH	C ARX–GARCHX	D ARXp–GARCHXp	E ARXp–GARCHX	F AR–GARCH
1	0.079 (0.002)	0.077 (0.002)	0.076 (0.001)	0.078 (0.003)	**0.070 (0.002)**	0.163 (0.003)
2	0.063 (0.001)	0.062 (0.001)	0.056 (0.001)	0.060 (0.002)	**0.054 (0.001)**	0.097 (0.002)
3	0.083 (0.001)	0.082 (0.001)	0.081 (0.001)	0.080 (0.001)	**0.076 (0.001)**	0.131 (0.002)
4	0.084 (0.000)	0.083 (0.000)	0.084 (0.000)	0.083 (0.002)	**0.075 (0.001)**	0.128 (0.001)

S5 WESAD Endpoint and Per-Participant Composition

The reference-based evaluation in Section 4.3 of the main text uses a binary window-level endpoint. The positive class is the set of windows annotated *stress* (WESAD label 2); the negative class pools all remaining retained windows, namely baseline (label 1), amusement (label 3), and transition/other (label 0). Transition and amusement windows are thus retained and treated as non-stress rather than discarded, so that the denominator reflects the full recording. After preprocessing (Section S8), 1,427 windows from 15 participants remained, of which 182 are stress and 1,245 non-stress. Table S3 gives the composition for each participant.

Table S3: Per-participant window and class counts in the WESAD analysis. "Non-stress" pools baseline, amusement, and transition/other windows.

Subject	Baseline	Stress	Amusement	Transition/other	Total	Stress (pos.)	Non-stress (neg.)
2	20	12	7	61	100	12	88
3	20	12	7	68	107	12	95
4	21	11	7	67	106	11	95
5	21	11	7	64	103	11	92
6	21	12	7	76	116	12	104
7	21	12	8	45	86	12	74
8	21	13	7	49	90	13	77
9	20	11	7	48	86	11	75
10	21	13	7	49	90	13	77
11	21	13	7	45	86	13	73
13	21	12	8	50	91	12	79
14	20	12	8	51	91	12	79
15	21	13	8	44	86	13	73
16	21	12	7	52	92	12	80
17	21	13	7	56	97	13	84
Total	311	182	109	825	1427	182	1245

Because the 1,427 windows are clustered within only 15 participants, a pooled DeLong test that treats windows as independent overstates precision. We therefore additionally quantify uncertainty with a *participant-clustered* bootstrap that resamples whole participants with replacement (3,000 resamples) and recomputes the pooled AUC on each resample. Re-scoring the saved model fit reproduces the AUCs reported in the main text (risk score 0.931, raw `HR_mean` 0.868) to within 0.005 (0.936 and 0.870). The participant-clustered 95% confidence intervals are [0.887, 0.972] for the risk score and [0.821, 0.918] for `HR_mean`. For the paired difference the interval is [0.031, 0.094], which excludes zero: the improvement of the variance-aware risk score over the raw heart rate is robust to within-participant clustering.

S6 Sensitivity to the Threshold Choice (Information-Leak Check)

The threshold $c_s(0.70)$ used in the main analysis is the 70th percentile of each participant's `HR_mean` over the *entire* recording. Although this uses no stress labels, it does include the high-

heart-rate windows that later form the positive class, which could lend a modest optimism to the reported discrimination. To assess this, we recomputed the exceedance score using thresholds set without such information—from an early per-participant burn-in window (the first 20% or 30% of each participant's windows) and, for reference, from that participant's baseline-annotated windows only—while holding the fitted conditional mean and variance fixed. Results are summarized in Table S4 and Figure S1.

Table S4: Discrimination of the annotated stress state (AUC) as a function of how the subject-specific threshold $c_s(0.70)$ is set. The conditional mean and variance are unchanged; only the threshold differs.

Rule for setting $c_s(0.70)$	AUC (stress vs. non-stress)
Full timeframe (main analysis)	0.936
Burn-in: first 20% of windows	0.876
Burn-in: first 30% of windows	0.886
Baseline-annotated windows only	0.870
Raw `HR_mean` baseline (reference)	0.870

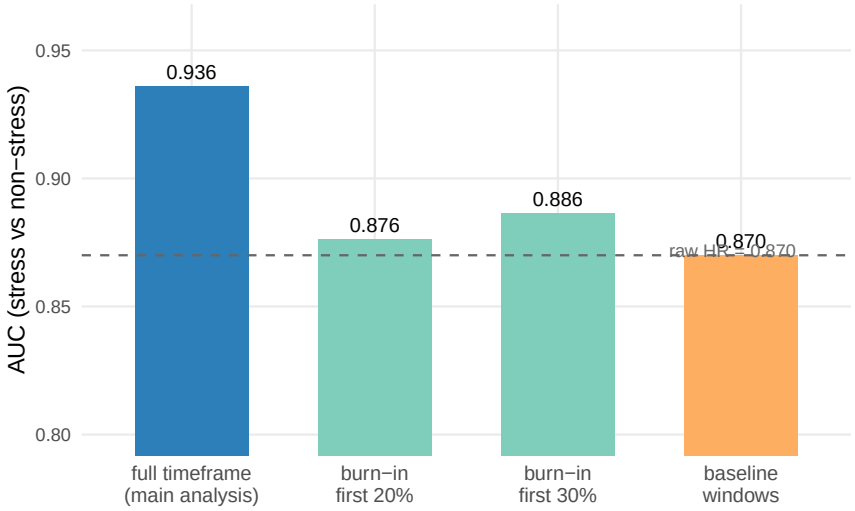


Figure S1: Risk-score AUC by the rule used to set the subject-specific threshold $c_s(0.70)$. A leak-free burn-in threshold reduces the AUC from 0.936 to approximately 0.88, which remains at or above the raw `HR_mean` baseline (dashed line).

Using a leak-free burn-in threshold reduces the AUC from 0.936 to approximately 0.88, still at or above the raw `HR_mean` baseline of 0.870. The temporal signal captured by the variance-aware score is therefore genuine, but the exact headline value depends on how the threshold is set; for prospective use the threshold should be fixed from an initial calibration period, as discussed in Section 5 of the main text.

S7 WESAD Variance-Covariate Ablation (Model B vs. Model D)

To isolate the contribution of covariate-driven volatility to the real-data results, we compare the full model (Model D, ARX$_p$–GARCHX$_p$: covariates in both the mean and the variance) with an otherwise identical model that removes covariates from the variance equation (Model B, ARX$_p$–GARCH). Both share the same penalized mean ($\lambda_\beta = 0.01$); they differ only in whether $x_{s,t}$ enters the conditional variance. We report discrimination of the annotated stress state (AUC) and, for the *actual* exceedance event $\{y_{s,t} > c_s(0.70)\}$ that the score is designed to predict, its discrimination (AUC) and calibration (Brier score). To keep the comparison controlled, both models are refit from scratch under identical settings in a single run; Model D here is the paper's main WESAD model, and its values differ from the headline AUC of 0.931 only through independent refitting.

Table S5: WESAD ablation. Discrimination of the annotated stress state and discrimination/calibration of the actual exceedance event $\{y > c_s(0.70)\}$ for the full model (D) and the no-variance-covariate model (B).

Model	AUC (stress)	AUC (event $y > c_s$)	Brier (event)
D: ARX$_p$–GARCHX$_p$ (X in variance)	0.921	0.927	0.085
B: ARX$_p$–GARCH (no X in variance)	0.934	0.929	0.084

Two conclusions follow. First, the variance-aware exceedance score—the central contribution—discriminates the annotated stress state far better than the raw heart rate (both models near 0.93 versus 0.868 for `HR_mean`) and is well calibrated for the exceedance event it is designed to predict (Brier $\approx$ 0.085, event rate 0.301). Second, whether covariates enter the *variance* equation makes little practical difference on WESAD: Models B and D differ by at most 0.013 in AUC and 0.001 in Brier, which is an order of magnitude smaller than the participant-level sampling uncertainty (the participant-clustered 95% interval for a single AUC spans roughly $\pm$0.045; Section S5), so the two are statistically indistinguishable here. This is consistent with the simulation study, where covariate-driven variance helps specifically when the volatility is covariate-structured (Scenarios 2–4): WESAD heart-rate volatility appears to be captured adequately by the GARCH dynamics alone, so the additional variance covariates neither help nor hurt materially. The exceedance score is therefore robust to this modeling choice, and Model D is retained as the general specification that also accommodates the covariate-driven-volatility

regimes demonstrated in simulation.

S8 Reproducibility Details for the WESAD Analysis

Features. Windowed features were computed per 60-s window (60-s hop) from the chest-worn RespiBAN streams. Moment- and correlation-based construction produced a larger candidate set, which was reduced to the 77 used for modeling by removing one member of each pair with absolute Pearson correlation above 0.90 (`caret::findCorrelation`). The complete list of the 77 retained features, grouped by sensing block, is given in Table S6.

Table S6: Complete list of the 77 chest-worn features used in the WESAD model, grouped by sensing block (counts in parentheses).

Block	Features
ECG / HRV (32)	HRV_ApEn, HRV_C1a, HRV_CD, HRV_CSI_Modified, HRV_CVI, HRV_Ca, HRV_DFA_alpha1, HRV_FuzzyEn, HRV_HF, HRV_HFD, HRV_HFn, HRV_HTI, HRV_KFD, HRV_LF, HRV_LFHF, HRV_LZC, HRV_LnHF, HRV_MCVNN, HRV_MinNN, HRV_PAS, HRV_PI, HRV_PSS, HRV_S, HRV_SD1SD2, HRV_SD2d, HRV_SDRMSSD, HRV_SI, HRV_SampEn, HRV_ShanEn, HRV_TINN, HRV_pNN20, HRV_pNN50
Respiration & BRV (9)	BRV_CV, BRV_RMSSD, Duty_cycle, ExpDur_mean, InspDur_mean, RSP_amp_mean, RSP_amp_std, RSP_rate_mean, RSP_rate_std
EDA (9)	EDA_kurt, EDA_mean, EDA_skew, SCL_CV, SCL_slope_cpm, SCL_std, SCR_amp_mean, SCR_area, SCR_count_pm
EMG (4)	EMG_MDF, EMG_Power, EMG_SpecEnt, EMG_ZCR
Accelerometry (9)	ACC_kurt, ACC_skew, ACC_std, ACCjerk_RMS, ACCmag_P90, ACCmag_RMS, ACCx_med, ACCy_med, ACCz_med
Skin temperature (4)	TEMP_detrended_std, TEMP_mean, TEMP_skew, TEMP_slope_cpm
Cross-channel correlations (10)	corr_ACC4_EMG4_4hz, corr_ACC4_TEMP4_4hz, corr_RSP4_ACC4_4hz, corr_RSP4_EMG4_4hz, corr_RSP4_SCL4_4hz, corr_RSP4_TEMP4_4hz, corr_SCL4_ACC4_4hz, corr_SCL4_EMG4_4hz, corr_SCL4_TEMP4_4hz, corr_TEMP4_EMG4_4hz

Windows, missingness, and the panel. A window was retained only if at least 80% of its samples were finite in every required channel; windows with any non-finite feature after construction were dropped and no imputation was used. The retained window count for each

participant is given in Table S3 (range 86–116; participants S1 and S12 are excluded following the WESAD convention). Because participants contribute different numbers of windows, the analyzed panel is *unbalanced*. The balanced-panel notation of the main text is expositional only: the estimator builds all y-, residual-, variance-, and covariate-lags *within* each participant (Algorithm 1), so unequal series lengths are handled directly, without truncation or padding, and only the shared parameters $(\theta, \beta, a, b, \gamma)$ are pooled across participants.

Tuning grid and temporal folds. Penalties were selected to preserve temporal ordering: each participant's windows were split into contiguous (non-shuffled) time-series folds, and validation losses were aggregated across participants (blockwise time-series cross-validation). In the simulations, λ_β was searched on a multiplicative grid around 5×10^{-3} and λ_γ on a logarithmically spaced grid refined around its minimizer. For the WESAD analysis the selected values were $\lambda_\beta = 1 \times 10^{-2}$ (mean) and $\lambda_\gamma = 1 \times 10^{-4}$ (variance), with fixed lag orders $(P, Q, L, R, M) = (1, 0, 1, 1, 0)$ and 20 outer iterations.

Software and code. Preprocessing and feature extraction used Python with NeuroKit2 (ECG/PPG/EDA/respiration features). Model fitting and evaluation used R with `glmnet` (penalized updates), `pROC` (ROC/AUC and DeLong test), and `randomForest` and `gbm` (supervised comparators); collinearity pruning used `caret`. Exact package versions and the code release (commit hash) used for the reported results are: R 4.5.3, with `glmnet` 5.0, `pROC` 1.19.0.1, `randomForest` 4.7.1.2, `gbm` 2.2.3, and `caret` 7.0.1; feature extraction used Python with NeuroKit2. The exact analysis code is released at the repositories given in the main-text Code Availability statement. The `varGuidTS` package and the replication scripts are available at the repositories listed in the main-text Code Availability statement.