

Article

SCOUT: Scenario-Oriented Context-Gated Forecasting with Explicit Lower-Bound Projection for Eligible Railway Coal OD Flows

Jianhao Zhu, Junhua Chen * and Ziying Ouyang

School of Traffic and Transportation, Beijing Jiaotong University, No. 3 Shangyuancun, Beijing 100044, China

* Correspondence: cjh@bjtu.edu.cn

Abstract

Heterogeneity in route activity and the need to clearly distinguish among observations, scenario assumptions, and operational requirements pose challenges for the short-horizon forecasting of railway coal origin–destination (OD) flows. SCOUT implements overlapping design objectives for longer-term tendencies, periodic variation, and short-term changes through 15 explicitly defined temporal and constructed-proxy reference components. Their construction provides temporal meaning, while a context-dependent simplex gate adapts their combination. Validation-based calibration is followed by separate scenario adjustment and route-level lower-bound projection. The case study concerns one railway system and 210 routes retained after training-period screening of inactive and rare-activation routes, covering 98.42% of held-out network flow. Across 171 held-out forecast origins under a purged chronological split, the lower-bound-compliant forecast reduces RMSE relative to Recent-4 by 6.34%, 3.21%, and 2.89% in the intermittent, regular active-flow, and backbone strata. The directional comparison with Recent-4 persists under the reported temporal-boundary and threshold checks, while the modern benchmarks and architectural diagnostics show cluster-dependent trade-offs rather than a uniform benefit from every module. Projection eliminates violations of the specified lower bounds, but does not establish network-capacity feasibility. For intermittent routes, positive-event coverage of the descriptive residual band is only 4.73%. The evidence supports a structurally organized forecasting representation and an auditable adjustment interface; transfer to other systems and downstream operational benefits remain to be evaluated.

Keywords: railway coal transportation; origin–destination forecasting; context-gated mixture; multi-horizon forecasting; scenario stress test; lower-bound projection



Academic Editor: Francesco Zirilli

Received: 25 August 2026

Revised: 2 October 2026

Accepted: 6 October 2026

Published: 10 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

1.1. Background and Motivation

Railway coal flows connect mines, industrial users, power plants, interchange yards, and ports. Accurate short-horizon origin–destination (OD) forecasts support train-plan preparation, loading and unloading coordination, corridor-capacity allocation, rolling-stock use, and priority energy supply. Coal demand and trade respond to electricity demand, fuel substitution, regional policy, and supply conditions [1]. Freight movements are further shaped by service-network design, intermodal opportunities, shipment-size decisions, transport-chain choice, and physical capacity [2–4]. An OD forecast is therefore both a statistical estimate and an input to network operations.

1.2. Problem Statement and Challenges

Three characteristics make this task materially different from forecasting a single dense series. First, route activity is strongly heterogeneous: many nominal OD cells are inactive, intermittent routes contain long zero spells and pulses, and a small backbone carries very large continuous volumes. Second, observed history, assumed future market inputs, and mandatory operating rules have different epistemic status. Combining them inside one opaque predictor makes it difficult to distinguish a data-driven estimate from a planner-specified stress test or a policy-enforced adjustment. Third, a model can appear favorable when route screening, temporal splits, percentage-error denominators, policy violations, or prediction intervals are defined ambiguously. A useful method must consequently control leakage, preserve interpretable decision interfaces, and expose the complete path from training state to published forecast. SCOUT's temporal design covers longer-term tendencies, periodic variation, and short-term jumps through inspectable reference forecasts. These are complementary design objectives, not mutually exclusive latent factors or an assumed causal decomposition.

1.3. Contributions

This study proposes SCOUT, a scenario-oriented context-gated forecasting architecture with explicit lower-bound projection. The study makes the following contributions.

1. It constructs route eligibility and three physically interpretable forecasting strata exclusively from the training period, separating intermittent, regular active-flow, and backbone routes.
2. It organizes forecasting information around overlapping objectives for longer-term tendencies, periodic variation, and short-term changes, implemented through 15 explicit reference components. The four functional groups index the implemented functions and ablation diagnostics rather than independent temporal submodels. A context-dependent simplex gate adaptively combines the components, whose temporal meaning comes from their construction.
3. It separates the statistical predictor, scenario operator, and lower-bound projection. This distinction permits stress testing without presenting assumed inputs as measured causal effects and guarantees only the stated route-level lower bounds.
4. It evaluates every eligible route under a chronological holdout and adds five rolling-origin folds, two shifted temporal boundaries, five random seeds, one-factor threshold sensitivity, component and architectural ablations, and moving-block bootstrap sensitivity.

The remainder of the paper reviews the relevant literature, develops the SCOUT model, specifies the training and inference algorithm, and presents a railway coal OD case study followed by discussion and conclusions.

2. Literature Review

2.1. Railway Freight and Coal OD Modeling

Freight OD analysis has long combined partial observations, assignment structure, and regional production–consumption relationships. Commodity-flow estimation, rail-freight assignment, and coal-specific OD reconstruction provide the direct domain foundation for the present study [5–7]. Recent railway-demand studies also demonstrate the explanatory value of economic, industrial, and operational variables [8]. This literature clarifies the network and commodity context, but much of it emphasizes matrix estimation, long-range demand, or assignment rather than transparent multi-step daily prediction under explicit scenario and operating rules.

2.2. OD Estimation and Dynamic Prediction

Classical OD inference includes entropy, likelihood, and generalized least-squares formulations [9–11]. Dynamic state-space methods model time-varying OD demand, while later neural formulations learn nonlinear temporal and network dependence [12–14]. Urban-rail studies show that recurrent, convolutional, and partial-observation structures can recover dynamic OD patterns [15,16]. However, passenger OD matrices are usually denser and more periodic than railway coal flows. Their methods do not directly resolve the combination of extreme route sparsity, order-of-magnitude scale differences, and route-level policy requirements considered here.

2.3. Spatiotemporal and Deep Traffic Forecasting

Short-term traffic forecasting has progressed from statistical models to recurrent, convolutional, attention-based, and graph neural networks [17–19]. DCRNN, STGCN, Graph WaveNet, GMAN, and AGCRN illustrate increasingly flexible representations of temporal dependence and network structure [20–24]; broader surveys document the resulting graph-forecasting design space [25]. Mixture-of-experts models provide another route to adaptive specialization. MoLE jointly trains linear forecasting experts and a router [26], whereas SCOUT mixes predefined statistical reference functions before separately applying scenario and lower-bound operators. This is a difference in model construction, not evidence that SCOUT outperforms MoLE. Contextual traffic forecasting also combines LSTM histories with weather and calendar features [27]; SCOUT's constructed profiles must not be treated as equivalent to such observed contextual measurements. For intermittent series, temporal-structure and modified TSB studies emphasize that zero spells and event timing require diagnostics different from dense-series forecasting [28,29]. Their expressiveness is valuable, but latent representations do not automatically reveal how recent level, seasonal recurrence, assumed future pressure, and an operating lower bound contribute to a particular forecast. For decision support, accuracy must be accompanied by an interface whose assumptions and adjustments can be inspected.

2.4. Structured and Constrained Forecasting

Forecast reconciliation explicitly distinguishes base statistical forecasts from outputs that satisfy aggregation, sign, or structural constraints. Optimal-combination, minimum-trace, and non-negative reconciliation methods establish this principle in hierarchical forecasting [30–33]. A priority-route minimum is a route-level analogue: policy context can inform the predictor, but a projection is still required when compliance must be guaranteed. At the same time, a projected value should not be reported as if it were an unconstrained statistical estimate.

2.5. Research Gap and Positioning of SCOUT

The reviewed streams leave a combined, rather than individually unprecedented, gap. Representative freight OD models supply network meaning but do not jointly provide short-horizon sparse-route learning, an explicit planner-controlled scenario map, and an auditable rule-compliance operator. Graph and attention models supply flexible dependence representations but usually embed mechanisms in latent states. Intermittent-demand methods address zero spells but not railway OD scenario authority or route-level policy rules. Reconciliation methods guarantee specified structural relations but do not prescribe how a heterogeneous railway OD base forecast should be learned. SCOUT addresses this intersection by using train-only route morphology, an auditable context-dependent mixture, and separate downstream scenario and lower-bound mappings. The contribution is not that any one ingredient is unprecedented, nor that component families identify unique

physical causes; it is the explicit composition and evaluation of these roles under a common leakage-controlled protocol. Within this composition, SCOUT specifies statistical reference functions with explicit temporal roles and learns their combination, rather than learning all expert forecasting functions jointly with the router. The construction-level distinction discussed in Section 2.3 clarifies the representation used here; it does not, by itself, establish superior accuracy over the cited methods.

Table 1 summarizes the positioning of these method streams.

Table 1. Positioning of representative method streams relative to the joint problem addressed here. A check mark denotes an explicit primary capability in the cited stream, not universal coverage by every paper. Dashes indicate that the stated joint interface is not the focus of the cited examples, rather than a systematic proof of absence across a whole literature.

Method Stream	Sparse/Heterogeneous Routes	Learned Predictor	Explicit Scenario Map	Guaranteed Stated Rule
Freight OD estimation/assignment	✓	–	–	–
Graph/attention traffic forecasting	partial	✓	–	–
Intermittent-demand forecasting	✓	✓	–	–
Forecast reconciliation	partial	–	–	✓
SCOUT	✓	✓	✓	✓

3. SCOUT: Scenario-Oriented Context-Gated Forecasting and Decision Operators

3.1. Notation and Forecasting Problem

Let $i = (o, d)$ index an OD route, t denote the last observed day, and $h \in \{1, \dots, H\}$ denote a direct forecast horizon. The observed non-negative flow is $y_{i,t}$, the lookback length is L , and $c(i)$ is the train-period forecasting stratum. Table 2 summarizes the principal notation.

Table 2. Principal symbols used in the SCOUT formulation.

Symbol	Meaning
$i = (o, d), t, h, k$	Route, forecast origin, direct horizon, and component indices
$y_{i,t}; L, H$	Observed daily flow; lookback and forecast-horizon lengths
$\mathbf{z}_{i,t,h}$	Standardized route–time–horizon context vector
$b_{i,t,h,k}$	Non-negative value of forecast component k
$\mathbf{g}_{c,h}; G_c(\cdot)$	Cluster–horizon global logits; context-gate network
$a_{i,t,h,k}; \gamma_c$	Simplex component weight; bounded gate-logit amplitude
$\hat{\mathbf{y}}^{raw}, \hat{\mathbf{y}}^0, \hat{\mathbf{y}}^{final}$	Neural mixture, calibrated forecast, and lower-bound-compliant forecast
$s_{c,h}, \tilde{s}_{i,h}$	Validation-selected cluster and route-shrunk scales
$\tilde{b}_{i,h}, \tilde{\zeta}_h$	Route-shrunk residual bias and selected bias strength
$\mathcal{S}_\delta, \mathcal{P}_L$	Scenario and route-level lower-bound operators
τ_0, τ_1, q_B	Activity-screening, regular-route, and backbone thresholds
$\rho_L, K_L, \mathcal{M}$	Policy-bound ratio, number of priority routes, and active months

Given the history $\mathbf{y}_{i,t-L+1:t}$, route descriptors $\mathbf{r}_i$, historical and known future inputs, and an available lower-bound reference, SCOUT defines the forecasts in Equations (1) and (2):

$$\hat{\mathbf{y}}_{i,t+1:t+H}^0 = \mathcal{G}_{c(i)}(\mathbf{y}, \mathbf{x}^{hist}, \mathbf{x}^{known}, \mathbf{r}_i, \mathbf{L}_i), \quad (1)$$

$$\hat{\mathbf{y}}_{i,t+1:t+H}^{final} = \mathcal{P}_{L_i}[\mathcal{S}_\delta(\hat{\mathbf{y}}_{i,t+1:t+H}^0)]. \quad (2)$$

The ordering is intentional: observations determine $\mathcal{G}$, a documented assumption determines $\mathcal{S}_\delta$, and an operating rule determines $\mathcal{P}_L$.

3.2. Modeling Rationale and Structural Component Bank

The bank operationalizes the temporal design in the Introduction: longer-window levels and slope references represent slower tendencies; weekday, lagged, and previous-year references represent recurrence; and last-value, short-window, momentum, and quantile references allow adaptation to recent changes. Here, longer-term is relative to the short forecast horizon and the available reference windows, not a claim to estimate a uniquely defined long-run trend. Several components serve more than one objective. The external-proxy component supplies a separate assumed-input adjustment. The following functional grouping is therefore an implementation and ablation index, not a replacement for the temporal design rationale. Four functional bookkeeping groups are used: seven recent-level components (Recent-4, Recent-7, Recent-14, Recent-28, Last Value, Recency Smooth, and Recent-75Q); five seasonal-reference components (Last-Year, Week Avg., Seasonal Momentum, Lag-7, and Lag-14); two trend/momentum components (Slope and Momentum); and one external-proxy adjustment. This grouping records implemented function rather than a uniquely identified physical or causal decomposition, and the set is not claimed to be the only valid 15-component design.

For example, if $\bar{y}_{i,t}^{(4)}$ and $\bar{y}_{i,t}^{(14)}$ are recent means, $\bar{y}_{i,t,h}^{wk}$ and $\bar{y}_{i,t,h}^{wk-}$ are current and preceding four-week references, and x_{t+h}^D is the known demand-pressure input, three parameterized components are given in Equations (3)–(5):

$$b_{i,t,h}^{mom} = \left[\bar{y}_{i,t}^{(4)} + h \Delta_{i,t}^{(4)} \right]_+, \quad (3)$$

$$b_{i,t,h}^{sm} = \left[\bar{y}_{i,t,h}^{wk} + \beta_{sm} (\bar{y}_{i,t,h}^{wk} - \bar{y}_{i,t,h}^{wk-}) \right]_+, \quad (4)$$

$$b_{i,t,h}^{ext} = \left[\bar{y}_{i,t}^{(14)} (1 + \beta_x x_{t+h}^D) \right]_+. \quad (5)$$

The four functional groups have overlapping temporal roles. The overlapping moving windows trade responsiveness against smoothing, while Last Value and the upper-quantile reference retain different responses to recent pulses. Seasonal references supply recurrence at different lags, and slope/momentum terms allow local extrapolation. These choices emphasize complementary temporal coverage and explicit component roles, without establishing the optimality of the 15-component bank. The recency-smooth component uses linearly increasing weights from r_{min} to r_{max} . The remaining components are direct moving, lagged, quantile, or slope references and contain no learned coefficients. Figure 1 provides a system-level view of the four-horizon component bank, the 65-dimensional context gate, validation-only affine calibration, and the downstream scenario and lower-bound stages; the exact implemented gate, stratum-specific fitting, residual bands, and parameterization are specified in the following subsections.

3.3. Leakage-Controlled Route Stratification

For the training-date set $\mathcal{T}_{tr}$, define route mean μ_i and non-zero ratio π_i using Equation (6):

$$\mu_i = |\mathcal{T}_{tr}|^{-1} \sum_{t \in \mathcal{T}_{tr}} y_{i,t}, \quad \pi_i = |\mathcal{T}_{tr}|^{-1} \sum_{t \in \mathcal{T}_{tr}} \mathbb{I}(y_{i,t} > 0). \quad (6)$$

Routes never active in training are excluded. Routes with $0 < \pi_i < \tau_0$ form a screened rare-activation group. Among the remaining routes, backbone Cluster 3 is assigned first when μ_i is at or above the q_B training quantile; non-backbone routes with $\tau_0 \leq \pi_i < \tau_1$ form intermittent Cluster 1, and those with $\pi_i \geq \tau_1$ form regular Cluster 2. Every statistic and threshold is estimated before validation and test, preventing future route activity or scale from influencing membership.

The baseline thresholds have operationally readable roles: τ_0 removes routes with too few training activations for a continuous volume model, τ_1 separates intermittent from regularly active service, and q_B isolates the high-volume tail. They are case-study assignments rather than universal constants. One-factor perturbations $\tau_0 \in \{0.005, 0.020\}$, $\tau_1 \in \{0.15, 0.25\}$, and $q_B \in \{0.92, 0.96\}$ are reported in Section 5.7 under both configuration-local and Base-common-route evaluation.

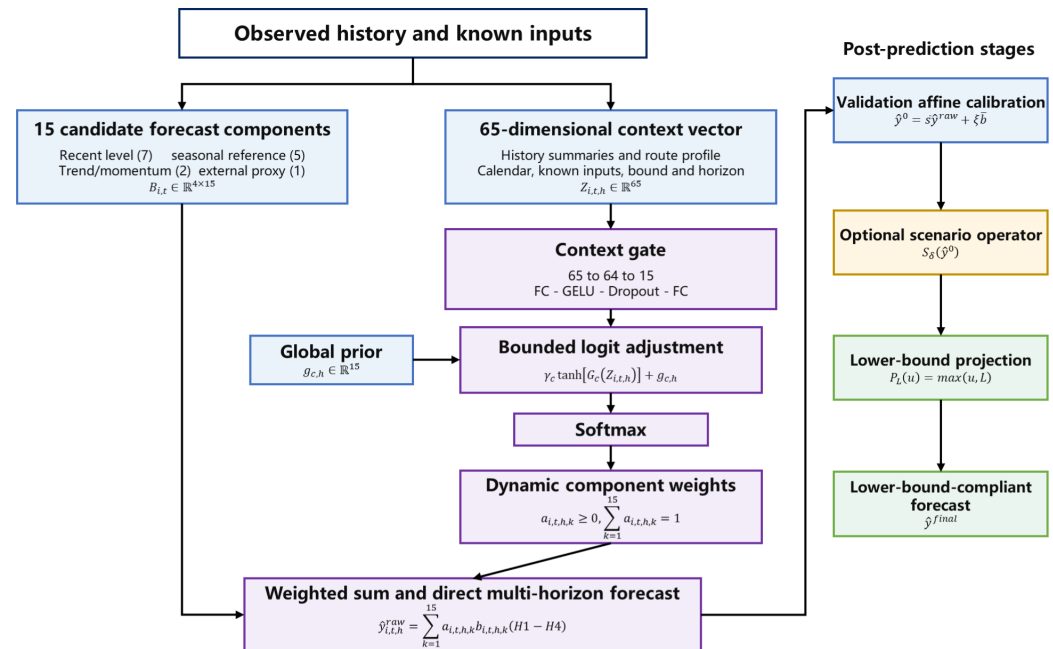


Figure 1. SCOUT context-gated forecasting and post-prediction decision pipeline. Observed history and constructed known-input proxies generate 15 candidate components over four direct horizons and a 65-dimensional context vector. The component bank comprises recent-level (7), seasonal-reference (5), trend/momentum (2), and external-proxy (1) functions. A 65–64–15 gate applies bounded adjustments to cluster–horizon global logits; softmax then yields simplex weights for the direct H1–H4 forecast. Validation-only affine calibration precedes the optional scenario operator and route-level lower-bound projection. Blue boxes identify component/context inputs, the global prior, and calibration; purple boxes show the learned mixture; yellow denotes the optional scenario operator; green denotes lower-bound projection and its output.

3.4. Context-Gated Simplex Mixture

The context $\mathbf{z}_{i,t,h} \in \mathbb{R}^{65}$ concatenates L log-transformed history values, 19 history-summary, train-only route-profile, and calendar terms, eight documented known inputs for the target day, and 10 horizon-reference terms. Training-set means and standard deviations normalize the context. A two-layer network with hidden width d_h and cluster-specific dropout probability p_c produces a bounded adjustment to cluster–horizon global logits, yielding the weights and forecast in Equations (7) and (8):

$$\mathbf{a}_{i,t,h} = \text{softmax}\left[\mathbf{g}_{c(i),h} + \gamma_c \tanh\{G_{c(i)}(\mathbf{z}_{i,t,h})\}\right], \quad (7)$$

$$\hat{y}_{i,t,h}^{\text{raw}} = \sum_{k=1}^K a_{i,t,h,k} b_{i,t,h,k}, \quad a_{i,t,h,k} \geq 0, \quad \sum_{k=1}^K a_{i,t,h,k} = 1. \quad (8)$$

The simplex combination preserves the units of the reference forecasts and permits their weighted contributions to be inspected. Their temporal meaning is specified by the component construction in Section 3.2; the gate implements adaptive combination rather than creating that meaning. The global logits are initialized from inverse training component

error and are then optimized jointly with the gate. Dropout is disabled during validation, calibration, and inference.

3.5. Neural Tensor Flow, Parameterization, and Initialization

For a cluster-specific full batch of N route–window samples, let $\mathbf{B} \in \mathbb{R}^{N \times H \times K}$ be the component tensor and $\mathbf{Z} \in \mathbb{R}^{N \times H \times d_z}$ the context tensor, where $(H, K, d_z) = (4, 15, 65)$. The sample and horizon dimensions of $\mathbf{Z}$ are merged to form an $(NH) \times 65$ matrix for application of the shared gate in Equations (9)–(11):

$$\mathbf{U} = \mathbf{Z}\mathbf{W}_1^T + \mathbf{b}_1, \quad \mathbf{W}_1 \in \mathbb{R}^{64 \times 65}, \quad \mathbf{b}_1 \in \mathbb{R}^{64}, \quad (9)$$

$$\mathbf{V} = \text{Dropout}(\text{GELU}(\mathbf{U}); p_c), \quad (10)$$

$$\mathbf{R} = \mathbf{V}\mathbf{W}_2^T + \mathbf{b}_2, \quad \mathbf{W}_2 \in \mathbb{R}^{15 \times 64}, \quad \mathbf{b}_2 \in \mathbb{R}^{15}. \quad (11)$$

After reshaping $\mathbf{R}$ to $N \times 4 \times 15$, the model computes $\mathbf{A} = \text{softmax}(\mathbf{g}_c + \gamma_c \tanh \mathbf{R})$ on the component axis and returns $\hat{\mathbf{Y}}^{raw} = \sum_k \mathbf{A} \odot \mathbf{B}$. The reported predictor consists of this context-gated mixture and does not add route embeddings, recurrent encoders, graph layers, attention blocks, or an alternative neural head.

Table 3 lists the trainable parameters of each cluster model.

Table 3. Trainable parameterization of one independently fitted cluster model. Horizon-specific states contain the same parameterization and are not additional jointly trained parameters.

Trainable Object	Shape	Parameter Count
Global logits $\mathbf{g}_c$	4×15	60
First-layer weights $\mathbf{W}_1$	64×65	4160
First-layer bias $\mathbf{b}_1$	64	64
Second-layer weights $\mathbf{W}_2$	15×64	960
Second-layer bias $\mathbf{b}_2$	15	15
Total per cluster	–	5259
Total across three independent clusters	–	15,777

The first-layer weights use Xavier-uniform initialization and all biases start at zero. The second-layer weights are then explicitly reset to zero, so the initial context adjustment is exactly zero. The global logits do not remain at zero. With $s_h = \max\{\text{median}_n |y_{n,h}|, 1\}$, their training-only component error and prior are defined in Equations (12) and (13):

$$e_{h,k} = N^{-1} \sum_n \left[\alpha_r \frac{|b_{n,h,k} - y_{n,h}|}{s_h} + \alpha_\ell |\log(1 + b_{n,h,k}) - \log(1 + y_{n,h})| \right], \quad (12)$$

$$g_{c,h,k}^{(0)} = \text{clip} \left(-\log \max(e_{h,k}, 10^{-4}) - K^{-1} \sum_j [-\log \max(e_{h,j}, 10^{-4})], -g_{\max}, g_{\max} \right). \quad (13)$$

Consequently, optimization starts from a transparent component prior and learns both its global logits and context-dependent departures. Context means and standard deviations are fitted only on the flattened training tensor; any standard deviation below 10^{-6} is replaced by one.

Dropout is active during optimization and disabled during validation, calibration, component analysis, and prediction. For output horizon h , inference restores the complete validation-selected state for that horizon—both $\mathbf{g}_c$ and the two-layer gate parameters—which performs the forward pass, retains column h , and concatenates the four retained columns. Validation-fitted scale, residual-bias, and interval quantities are applied only after this neural forward pass and do not alter the trained parameters.

3.6. Validation Calibration and Residual Bands

Let $s_{c,h}$ be a cluster scale selected on validation data and $\tilde{s}_{i,h}$ a route candidate. With n_i^{val} route-validation samples and shrinkage constant κ_s , the shrunk scale is given by Equation (14):

$$\bar{s}_{i,h} = (1 - \omega_i)s_{c,h} + \omega_i\tilde{s}_{i,h}, \quad \omega_i = \frac{n_i^{val}}{n_i^{val} + \kappa_s}. \quad (14)$$

Route mean residuals are capped at ρ_b times the median absolute scaled forecast and shrunk toward the cluster residual mean using $v_i = n_i^{val} / (n_i^{val} + \kappa_b)$. A horizon-specific strength ζ_h gives the calibrated forecast in Equation (15):

$$\hat{y}_{i,t,h}^0 = \left[\bar{s}_{i,h}\hat{y}_{i,t,h}^{raw} + \zeta_h\bar{b}_{i,h} \right]_+. \quad (15)$$

The q_{lo} and q_{hi} validation-residual quantiles define marginal descriptive bands around $\hat{y}^0$. They are not presented as conditional or conformal guarantees.

3.7. Scenario and Route-Level Lower-Bound Operators

A market scenario uses a horizon-dependent multiplier $m_h = m_1 + (h - 1)(m_H - m_1) / (H - 1)$. The illustrative supply-security bound is defined in Equation (16):

$$B_{i,t} = \max \left\{ \max_h L_{i,t,h}, \rho_F \max_h \hat{y}_{i,t,h}^0, \rho_R \text{mean}(y_{i,t-W_S+1:t}) \right\}. \quad (16)$$

The combined scenario applies m_h first and then projects onto $[B_{i,t}, \infty)$. These quantities are controlled stress-test inputs rather than estimated elasticities.

For standard evaluation, the first K_L routes within each stratum are selected by training-period mean. For a selected route, the standard rule is activated for an entire four-day forecast window when the month of its first target day, $t + 1$, belongs to $\mathcal{M}$. The history-based bound is then the same at all four horizons, as specified in Equation (17):

$$L_{i,t,h} = \rho_L \max \{ \text{mean}(y_{i,t-W_L+1:t}), Q_{q_L}(y_{i,t-W_L+1:t}) \}. \quad (17)$$

Otherwise, $L_{i,t,h} = 0$ for the entire window. The activation condition is based on month($t + 1$), not separately on month($t + h$). All observed-flow references end at the forecast origin t ; earlier held-out observations can enter later forecasts only after they have been observed. The settings $K_L = 3$, $\rho_L = 0.58$, $W_L = 14$, $q_L = 0.75$, and the active-month set are fixed illustrative case-study assignments, not universal operating rules. Feasibility projection is instantiated here as Euclidean projection onto the specified route-level lower-bound set. Its output is therefore termed a lower-bound-compliant forecast. With $\mathcal{C}_L = \{\mathbf{z} : z_h \geq L_h\}$, the projection is given by Equation (18):

$$\mathcal{P}_L(\mathbf{u}) = \arg \min_{\mathbf{z} \in \mathcal{C}_L} \|\mathbf{z} - \mathbf{u}\|_2^2 = \max(\mathbf{u}, \mathbf{L}). \quad (18)$$

The projection is unique, idempotent, and non-expansive, and makes the minimum Euclidean adjustment required by the stated route-wise lower bound. It does not enforce shared corridor, terminal, or rolling-stock capacities.

4. Training and Inference Algorithm

4.1. Chronological Learning Protocol

The full pipeline separates parameter estimation, model selection, and final assessment. Route profiles, strata, component-prior initialization, context normalization, and gradient updates use training dates only. Validation dates determine early stopping, one

validation-selected parameter state per horizon, affine calibration, residual quantiles, and any comparator hyperparameter or blend. Test observations are reserved for final evaluation. This protocol also fixes the meaning of a forecast origin: it is the last observed day, while the forecast-window start is the first target day. Figure 2 summarizes this chronological separation and the distinct authority of estimation, decision, and held-out evaluation operations.

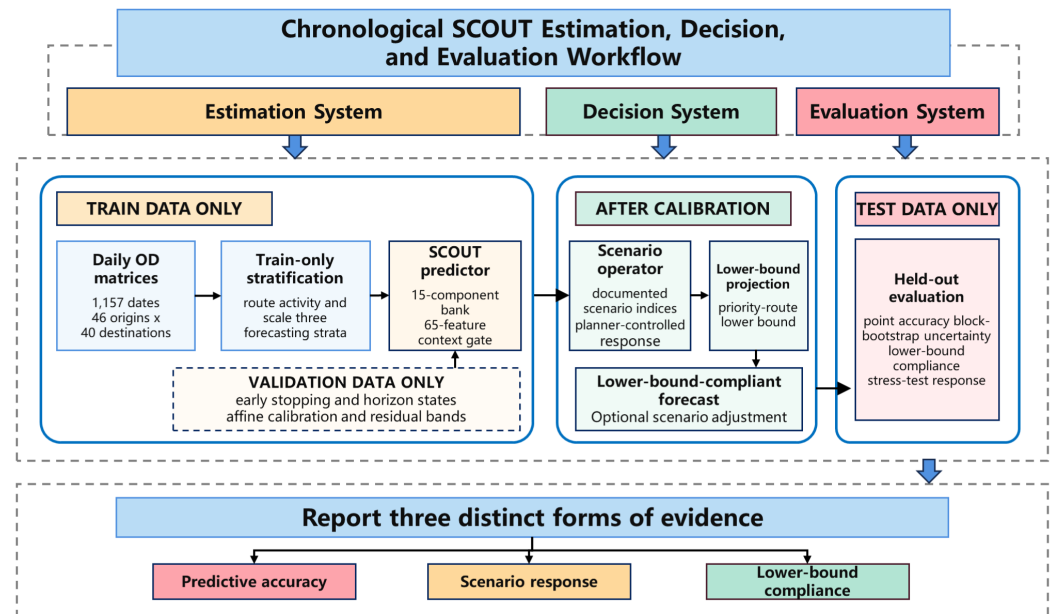


Figure 2. Chronological separation of SCOUT estimation, decision operations, and held-out evaluation. The 1157 daily OD matrices contain 46 origins and 40 destinations. Route stratification and predictor fitting use training data, while early stopping, horizon-state selection, affine calibration, and residual bands use validation data only. The documented scenario operator and route-level lower-bound projection are applied after calibration; held-out test data support predictive-accuracy, scenario-response, and rule-compliance evidence.

4.2. Objective Function and Optimization

Let ϵ_+ define a positive target, ω_+ its additional weight, $!^H$ the horizon weights, and s_c the cluster root-mean-square training target. The optimized objective is given in Equation (19):

$$\mathcal{L}(\theta) = \frac{1}{Z} \sum_{i,t,h} w_{i,t,h}^+ \omega_h^H \left[\lambda_r \frac{|\hat{y}_{i,t,h}^{raw} - y_{i,t,h}|}{\max(y_{i,t,h}, 1)} + \lambda_\ell \text{Huber}(\log(1 + \hat{y}_{i,t,h}^{raw}) - \log(1 + y_{i,t,h})) \right. \\ \left. + \lambda_s \left(\frac{\hat{y}_{i,t,h}^{raw} - y_{i,t,h}}{s_c} \right)^2 \right], \quad (19)$$

where $w^+ = \omega_+$ when $y > \epsilon_+$ and $w^+ = 1$ otherwise, and $Z = \sum_{i,t,h} w_{i,t,h}^+ \omega_h^H$ is the total batch weight. AdamW jointly updates the global logits and gate parameters. The gate learning rate is ρ_g times the cluster logit rate, and gradient norm is limited by G_{max} .

The loss weights were fixed as a design compromise before held-out evaluation: the relative and log-Huber terms reduce domination by route scale and the squared term preserves sensitivity to large absolute misses. They are not physical constants. Local alternative weight triples produced cluster RMSE changes of at most approximately 0.01% in the local sensitivity diagnostic (Appendix B), indicating numerical insensitivity near the reported setting; transfer to another railway system should nevertheless revalidate the loss, horizon, and positive-target weights.

The validation selection score is expressed symbolically in Equation (20):

$$J_{val} = \eta_R \text{RMSE}_{norm} + \eta_P \text{MAPE}_{norm} + \eta_A \text{MAE}_{norm}, \quad (20)$$

where each metric is normalized by the Recent-4 validation reference. The Recent-4 validation reference is stored independently of the candidate component tensor and is retained unchanged in every component-family diagnostic. Removing or reordering candidate outputs therefore does not alter the reference used to normalize the validation selection score. This reference is used for model selection only and is not reintroduced as a candidate in an ablation that removes it. Early stopping retains a minimum- J_{val} parameter state for every direct horizon. Test horizon h is generated by its corresponding validation-selected state.

4.3. Validation-Only Hyperparameter Refinement

Hyperparameter refinement compares the base gate with two alternatives using validation data only: a balanced gate $(\gamma_c, \rho_{g,c}, p_c) = (0.35, 0.10, 0.03)$ and a higher-capacity gate $(0.50, 0.20, 0.03)$. Candidate fits use at most 120 epochs and a patience of 35 epochs. An alternative is selected only when it reduces J_{val} by at least 2×10^{-4} . Under the purged split, C1 and C3 retain the initial setting, whereas C2 selects the higher-capacity gate, reducing its validation score from 0.939099 to 0.938725. All settings are fixed before held-out evaluation, and test outcomes do not enter the comparison.

4.4. Training and Inference Procedure

Algorithm 1 summarizes the training, calibration, and inference procedure.

Algorithm 1: Train, calibrate, and apply SCOUT

1. Split dates chronologically; purge every sample whose complete four-day target window crosses a boundary; compute μ_i , π_i , eligibility, and $c(i)$ from training targets only.
 2. Materialize K component forecasts and the 65-dimensional contexts; fit context normalization on training samples.
 3. Initialize $\mathbf{g}_{c,h}$ from the inverse weighted component error, using relative- and log-error coefficients α_r and α_ℓ and clipping bound g_{max} .
 4. For each epoch, perform one full-batch AdamW update of $\theta = \{\mathbf{g}, G\}$, evaluate validation J_{val} , save global and horizon-specific improved states, and stop after P non-improving epochs.
 5. Restore the selected horizon states. On validation data, select cluster and route-shrunk scales, capped residual biases, bias strengths, and residual quantiles.
 6. For each held-out sample, compute $\hat{y}^{raw}$ from the corresponding validation-selected model parameters and then $\hat{y}^0$ from validation calibration. Retain this value as the unconstrained point forecast before any policy projection.
 7. When requested, apply the documented scenario operator and then $\mathcal{P}_L$. Report unconstrained and lower-bound-compliant forecasts separately.
-

4.5. Parameter Assignment and Implementation Settings

Tables 4 and 5 assign every substantive coefficient appearing above. These case-study values were fixed before held-out evaluation and are not inferred from test outcomes. They document this case study rather than universal recommendations: calendar and component windows encode the present forecasting task, optimization settings are implementation choices, calibration quantities are validation-derived, and scenario/lower-bound values are planner-defined examples that require local re-estimation or authorization when transferred.

Cluster logit learning rates are denoted α_c . The gate rate is $\rho_{g,c}\alpha_c$, while γ_c , p_c , $E_{max,c}$, and P_c denote the gate amplitude, dropout, epoch limit, and patience for stratum c . Table 6 gives the final validation-selected assignments.

The primary run uses random seed $s_0 = 42$. Each epoch uses one full-batch update because a route-window sample contains all four direct horizons. PyTorch is used for optimization, automatic differentiation, and state restoration. The seed controls parameter initialization and dropout sampling. Training stability is assessed separately over seeds 11, 23, 42, 67, and 101.

Table 4. Cluster-independent SCOUT model, loss, and optimization assignments.

Symbol	Assigned Value	Role
L, H, K, d_z, d_h	28, 4, 15, 65, 64	Lookback, horizons, components, context, and hidden width
$\alpha_r, \alpha_\ell, g_{max}$	0.65, 0.35, 1.80	Component-prior error weights and logit clip
$r_{min}, r_{max}, \beta_{sm}, \beta_x$	0.35, 1.00, 0.45, 0.08	Parameterized component coefficients
ϵ_+, ω_+	0.10, 4.0	Positive-target threshold and loss weight
$\lambda_r, \lambda_\ell, \lambda_s$	0.58, 0.32, 0.10	Relative, log-Huber, and squared loss weights
$!^H$	(0.85, 1.15, 1.70, 2.45)	Horizon loss weights H1–H4
η_R, η_P, η_A	0.50, 0.35, 0.15	Validation score weights
$G_{max}, \Delta_{min}, \lambda_{wd}$	5.0, 10^{-5} , 5×10^{-5}	Gradient clip, minimum improvement, and weight decay

Table 5. Stratification, calibration, scenario, and lower-bound parameter assignments.

Symbol	Assigned Value	Role
τ_0, τ_1, q_B	0.01, 0.20, 0.94	Route activity and backbone thresholds
$[s_g^-, s_g^+], N_g$	[0.96, 1.04], 9	Cluster calibration grid
$\delta_s, [s_r^-, s_r^+], N_r$	0.02, [0.92, 1.08], 13	Route scale band, bounds, and grid size
$\kappa_s, \rho_b, \kappa_b$	128, 0.12, 32	Scale shrinkage, bias cap, and bias shrinkage
Ξ	{0, 0.25, 0.50, 0.75, 1}	Candidate bias strengths ζ_h
q_{lo}, q_{hi}	0.10, 0.90	Validation-residual quantiles
m_1, m_H	0.97, 0.90	Market-shock endpoint multipliers
ρ_F, ρ_R, W_S	1.06, 1.03, 7	Scenario-bound forecast and recent-history references
K_L, ρ_L, W_L, q_L	3, 0.58, 14, 0.75	Priority routes and standard lower-bound reference
$\mathcal{M}$	{11, 12, 1, 2, 3}	Months with an active standard policy rule

Table 6. Cluster-specific optimization settings and validation-only selection outcome.

c	α_c	γ_c	$\rho_{g,c}$	p_c	$E_{max,c}$	P_c	Selected Setting
1	0.015	0.25	0.05	0.05	240	45	base gate
2	0.015	0.50	0.20	0.03	120	35	higher-capacity gate
3	0.030	0.25	0.05	0.05	240	45	base gate

Table 7 reports the training and validation sample counts and the selected epochs.

Table 7. Training samples and validation-selected epochs. H1–H4 are the states used for their respective test horizons.

Cluster	Routes	Train Samples	Val. Samples	α_c	H1	H2	H3	H4
1	126	52,038	21,546	0.015	13	7	1	6
2	71	29,323	12,141	0.015	13	8	2	1
3	13	5369	2223	0.030	17	1	8	1

5. Numerical Case Study

5.1. Case Description and Data

The case study contains 1157 consecutive daily railway coal-weight matrices from 1 November 2021 to 31 December 2024. The matrices contain 46 origins and 40 destinations,

yielding 1840 nominal OD cells. Across the full period, 379 cells are positive at least once; 345 are positive during training. Total observed weight is 1,465,364,505 in the provider-specific source unit. The available metadata do not support a defensible conversion to tonnes, wagons, trainloads, cost, or capacity, so no such conversion is applied in the analysis or figures. The first 70% of dates, ending 18 January 2024, define route eligibility and clustering. A forecast sample is assigned to a split only when its complete four-day target window lies inside that split. Training target windows therefore end no later than 18 January; validation windows start from 19 January through 7 July and end no later than 10 July; and windows starting 11 July–28 December are held out for testing, with targets extending through 31 December. The three forecast-window starts crossing each boundary are purged, so no route–target-date label is shared by training, validation, or test. This gives 413 training, 171 validation, and 171 test forecast windows per route; held-out forecast origins run from 10 July–27 December 2024. At each successive origin, lagged observations already available by that origin may include earlier validation or test dates; this is rolling forecasting with frozen fitted parameters, not forecasting the whole test period using only the initial training history. No target later than the current origin is available to that forecast.

Complete timestamped coal-market, energy, weather, and congestion measurements were not available. Deterministic proxy indices were therefore constructed for international and domestic coal prices, import pressure, power demand, heating degree, policy priority, port congestion, and composite demand pressure, and were supplied to the model as assumed known inputs. They are deterministic calendar profiles, not observations. In particular, `coal_import_pressure_index` is not computed from, smoothed from, or fitted to Gangqimaodu-origin OD flows. The indices make the experimental interface executable but do not support causal attribution or estimation of coal-price elasticity.

For reproducibility, let $d = 0, \dots, 1156$ be the day index, m the calendar month, and $u = d/1156$. Define $w = \max\{0, \cos[2\pi(m-1)/12]\}$, $s = \max\{0, \cos[2\pi(m-7)/12]\}$, $r = \sin(2\pi d/7)$, and $a = \sin(2\pi d/365.25)$. Before rounding to three decimals, the seven elementary proxies are defined in Equations (21)–(26):

$$x^{int} = 112 + 12 \sin \frac{2\pi(d-50)}{365.25} + 7 \sin \frac{2\pi d}{180} - 9u, \quad (21)$$

$$x^{dom} = 104 + 9 \sin \frac{2\pi(d+22)}{365.25} + 5w - 3u, \quad (22)$$

$$x^{imp} = 96 + 8 \sin \frac{2\pi(d+90)}{365.25} + 14 \exp[-((m-8.5)/2.2)^2], \quad (23)$$

$$x^{pow} = 100 + 8w + 7s + 4u + 1.5r, \quad x^{heat} = 100w, \quad (24)$$

$$x^{pol} = 40 + 55\mathbb{I}(m \in \{11, 12, 1, 2, 3\}) + 10 \sin \frac{2\pi d}{365.25}, \quad (25)$$

$$x^{port} = 50 + 10 \sin \frac{2\pi(d+40)}{90} + 6 \max(0, a). \quad (26)$$

The seven elementary profiles are first rounded to three decimals. Standardization $z(\cdot)$ then uses their means and population standard deviations (ddof = 0) over the fixed 1157-day calendar domain, after which the eighth proxy is given by Equation (27):

$$\begin{aligned} x^D &= 0.30z(x^{pow}) + 0.28z(x^{heat}) + 0.16z(x^{dom}) + 0.18z(x^{pol}) \\ &\quad - 0.22z(x^{imp}) - 0.10z(x^{int}) - 0.06z(x^{port}). \end{aligned} \quad (27)$$

The composite is also rounded to three decimals. This calendar-domain normalization defines the synthetic scenario profiles; it does not use OD targets and is distinct from the subsequent context normalization fitted on training samples only. All temporal experiments reuse this documented calendar domain rather than re-estimating proxy profiles from their

test outcomes. Extending the deployment calendar would require an explicit reference-domain convention and is not validated here. Because these profiles depend only on the calendar and fixed formulas, their target-date values are available at each forecast origin by construction. This availability property is not evidence that they measure the named market quantities.

Table 8 summarizes the training-period route strata.

Table 8. Training-period route strata. Mean flow and non-zero ratio are computed before validation and test.

Cluster	Operational Morphology	Routes	Train-Period Mean Flow	Mean Non-Zero Ratio
0	Screened rare-activation routes	135	16.80	0.0038
1	Intermittent pulse routes	126	329.49	0.0597
2	Regular active-flow routes	71	6993.24	0.6067
3	Backbone routes	13	55,760.96	0.9969

The train-period q_B threshold corresponding to Cluster 3 is 27,041.05. The 210 eligible routes account for 99.82%, 99.31%, and 98.42% of total observed network flow in the training, validation, and test phases, respectively. During the test phase, 42 of 135 screened rare routes activate and contribute 1.22% of flow; 22 of 1495 training-never-active routes activate and contribute 0.36%. Thus, the main evaluation covers most volume but does not evaluate network-wide route activation.

5.2. Experimental Design and Comparators

Four evaluation variants distinguish predictive and rule-compliance effects: Recent-4; Recent-4 followed by lower-bound projection; calibrated SCOUT before projection; and calibrated SCOUT followed by lower-bound projection. All reported eligible test routes are included; displayed routes are selected only by training-period criteria.

The external pre-projection suite uses identical routes, forecast origins, horizons, and chronological boundaries, with the method-specific inputs described below. It includes last value, lag-7 seasonal naive, Recent-4, direct Ridge, quantile gradient boosting (GBDT), a direct GRU, and SCOUT. Let $\mathcal{A}_R$ denote the Ridge penalties and $\mathcal{B}_R$ its Recent-4 blend grid; they are assigned $\mathcal{A}_R = \{0.1, 1, 10, 100\}$ and $\mathcal{B}_R = \{0, 0.125, \dots, 1\}$. Ridge and GRU blend selection uses J_{val} . GBDT uses quantiles $\mathcal{Q}_G = \{0.1, 0.5, 0.9\}$, iterations $E_G = 100$, maximum leaves $M_G = 24$, learning rate $\alpha_G = 0.045$, and regularization $\lambda_G = 0.05$. Its median blend is selected from an $N_{BG} = 11$ point grid by $J_G = \text{RMSE}_{norm} + \lambda_{GP} \text{MAPE}_{norm}$ with $\lambda_{GP} = 0.35$. The direct GRU uses hidden width $d_G = 32$, head width $d_{head} = 64$, dropout $p_G = 0.05$, AdamW rate $\alpha_{GRU} = 0.003$, weight decay $\lambda_{GRU} = 10^{-4}$, batch size $B_G = 1024$, maximum epochs $E_{GRU} = 36$, and patience $P_{GRU} = 8$. Every fitted comparator and blend weight is selected on validation data only. Both GRU and SCOUT are implemented in PyTorch and follow the same chronological model-selection protocol.

Two GRU protocols must be distinguished. The original benchmark uses a 36-epoch maximum; the additional comparison runs use a 48-epoch maximum. Both use patience 8, hidden width 32, head width 64, and seed 42 for single-seed comparisons. They are separate fitted runs, not interchangeable estimates of one checkpoint. The original paired comparator intervals below refer only to the 36-epoch run. Different stopping budgets and stochastic optimization can produce different selected predictions; no exact attribution of their difference to one factor is claimed.

The Patch Transformer is a custom patch-based comparator, not an official PatchTST replication: it embeds four non-overlapping seven-day history patches into width 64, uses two four-head encoder layers (feed-forward width 128, dropout 0.1), and combines the

flattened encoding with static/context features in a 128-unit direct-output head. GRU and Patch Transformer receive 28 log-history values plus history summaries, route/calendar descriptors, and the four horizons' proxy/reference inputs. Their training-only scaling and validation-only Recent-4 blending follow the same additional-comparison protocol; maximum epochs are 48, patience 8, and batch size 1024, with AdamW rates 0.003 and 0.001, respectively. These are declared method-specific budgets, not equal computation or an exhaustive architecture search.

The two-stage intermittent comparator fits a histogram gradient-boosting activation classifier and a positive-volume regressor on log-transformed targets separately at each horizon. Both use 100 iterations, 24 maximum leaves, learning rate 0.05, and L2 regularization 0.05. A validation-selected probability threshold from 0.20 to 0.80 in steps of 0.05 determines whether the positive-volume prediction is emitted; this is a thresholded output, not probability times conditional mean, and it is not an implementation of TSB. The OD-graph comparator is a validation-tuned diffusion reference over routes sharing an origin or destination, not a graph neural network; peer forecasts use the median of neighboring Recent-4/Recent-14 ratios rescaled by the focal route's Recent-14 level. For each horizon, the relation (shared origin, destination, or either) and Recent-4 blend coefficient (0 to 0.5 in steps of 0.05) are selected on validation data. Algorithm 2 summarizes the common evaluation workflow; the model-specific inputs and tuning budgets are specified above.

Algorithm 2: Common comparator workflow

1. Fix the same eligible routes, four-day targets, and purged chronological split used by SCOUT; fit every scaler or route statistic on training data only.
 2. Supply persistence methods only their defining lags; supply Ridge, GBDT, GRU, Patch Transformer, two-stage intermittent, and OD-graph diffusion implementations the documented historical inputs and their method-specific structure. No comparator receives targets later than the current forecast origin or unavailable future observations; earlier held-out observations may enter lagged inputs once observed.
 3. Fit each candidate configuration on training data and select regularization, blend, stopping epoch, or method configuration on validation data under its stated score and budget.
 4. Freeze the selected state, predict the identical route–origin–horizon keys, and report pre-projection point metrics. Apply the same lower-bound operator where specified, and explicitly label post-projection results.
-

5.3. Evaluation and Statistical Inference

Point accuracy is evaluated with MAE, RMSE, positive-target MAPE, WAPE, and sMAPE. Positive-target MAPE uses cells with $y > 0.1$ and an absolute numerator. This fixed numerical cutoff excludes zero and very small denominators; it is not a physical operating threshold. Policy violation is reported conditionally over cells with $L_{i,t,h} > 0$; the active-cell share is reported separately. Quantile outputs use pinball loss and empirical coverage. Paired uncertainty uses 2000 seven-day moving-block bootstrap replicates over forecast-origin dates, retaining route and horizon alignment within each sampled date. All methods use the same chronological boundaries, forecast origins, eligible routes, and horizon alignment. Reported statistics are computed from the corresponding validation-selected parameter states, with scenario operations applied before lower-bound projection.

5.4. Route Heterogeneity

Of 1840 nominal OD cells, 1495 were never active during training, and 135 additional routes were screened as rare-activation cases. The remaining 210 routes comprise 126 intermittent, 71 regular active-flow, and 13 backbone routes. Their mean train-period flows are 329.49, 6993.24, and 55,760.96, respectively, while mean non-zero ratios are 0.0597, 0.6067, and 0.9969. The resulting scale and activity differences justify cluster-specific optimization and metric interpretation: C1 is primarily an activation-and-volume problem, whereas C3 is a nearly continuous large-volume problem.

Figure 3 illustrates these differences in route activity and scale.

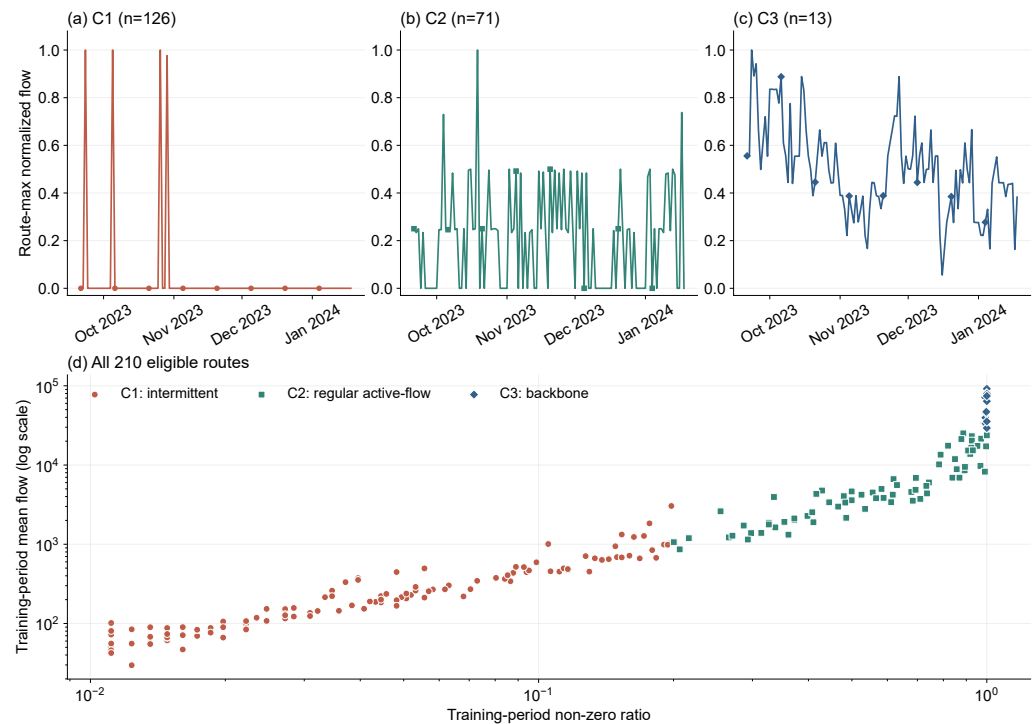


Figure 3. Distinct training-period route morphologies. Panels (a–c) show the final 120 training days for representative C1–C3 routes selected by distance to their stratum medians; each trace is normalized by its own maximum to expose temporal shape. Panel (d) includes all 210 eligible routes using unnormalized training statistics. Selection and axes do not use validation or test errors.

The shapes imply different physical tasks. C1 consists of long zero spells and isolated loading pulses, so a useful model must first anticipate activation and then estimate positive volume. C2 represents recurring but variable service flows, for which level, weekly recurrence, and short-run adjustment all matter. C3 is nearly continuously active and carries an order-of-magnitude larger mean, consistent with a trunk or backbone function whose main risk is a large absolute deviation. These interpretations follow directly from activity frequency and scale; they do not assign unobserved causes to individual routes.

5.5. Held-Out Accuracy and Lower-Bound Compliance

Table 9 reports all four evaluation variants. The lower-bound-compliant SCOUT forecast reduces RMSE from 1849.58 to 1732.25 in C1, from 5217.19 to 5049.90 in C2, and from 14938.40 to 14506.00 in C3. The corresponding reductions are 6.34%, 3.21%, and 2.89%. Positive-target MAPE improves by 2.03 and 3.18 percentage points in C1 and C2, but worsens by 0.30 points in C3. C2 WAPE is 0.13 points higher than Recent-4, C1 is essentially unchanged, and C3 improves. The principal cross-cluster finding is therefore control of large absolute errors, not universal dominance on every metric.

The projection has an operational rather than uniformly accuracy-improving role. Relative to SCOUT, SCOUT + projection slightly lowers C1 RMSE, slightly raises C2 MAE and RMSE, and changes C3 point errors negligibly. It nevertheless moves conditional violation from 13.58%, 6.68%, and 0.14% to zero. Policy-active cells account for only 0.54%, 0.96%, and 7.83% of all C1–C3 test cells, respectively; the conditional denominator is therefore essential. Of 464, 464, and 696 policy-active cells, the SCOUT projection changes 63, 31, and 1 cells, with mean adjustments among changed cells of 1490.64, 4361.38, and 113.55 source units. This is route-level rule compliance, not proof of corridor, terminal, fleet, or flow-conservation feasibility.

Table 9. Held-out test metrics. MAE and RMSE are in source weight units. Conditional violation uses active lower-bound cells only; active share is the percentage of all test cells with a positive lower bound.

Cluster	Variant	MAE	RMSE	MAPE (%)	WAPE (%)	Cond. Violation (%)	Active Share (%)
1	Recent-4	623.38	1849.58	63.64	101.76	30.17	0.54
1	Recent-4 + projection	622.24	1841.36	63.53	101.58	0.00	0.54
1	SCOUT	623.45	1733.89	61.61	101.77	13.58	0.54
1	SCOUT + projection	623.36	1732.25	61.61	101.76	0.00	0.54
2	Recent-4	3124.11	5217.19	46.66	48.37	24.14	0.96
2	Recent-4 + projection	3127.14	5207.15	46.74	48.41	0.00	0.96
2	SCOUT	3131.03	5047.18	43.47	48.47	6.68	0.96
2	SCOUT + projection	3132.20	5049.90	43.48	48.49	0.00	0.96
3	Recent-4	10,773.35	14,938.40	28.09	20.37	6.90	7.83
3	Recent-4 + projection	10,762.87	14,921.27	28.07	20.35	0.00	7.83
3	SCOUT	10,484.30	14,505.98	28.39	19.82	0.14	7.83
3	SCOUT + projection	10,484.31	14,506.00	28.39	19.82	0.00	7.83

5.6. External Benchmark Comparison

Table 10 compares pre-projection forecasts; SCOUT here is already affine-calibrated, not the uncalibrated neural mixture. SCOUT has the lowest RMSE and positive-target MAPE in C1 and C2. Its C1 RMSE is 4.27% below GBDT and 3.59% below GRU; the paired intervals are [3.00%, 5.50%] and [2.43%, 4.73%]. In C2, the RMSE gains are 4.51% over GBDT and 1.51% over GRU, while the GRU interval includes zero at [−0.54%, 4.01%]. C3 remains a near tie: SCOUT RMSE is 0.69% higher than GRU and the interval [−2.38%, 1.52%] includes zero, while SCOUT has lower MAE, MAPE, and WAPE than GRU. Relative to GBDT in C3, SCOUT's 0.53% RMSE advantage is also uncertain. The expanded benchmark therefore supports strong C1 evidence, a C2 large-error advantage with metric dependence, and no statistically resolved RMSE difference from the strongest comparator in C3 under this analysis. A confidence interval containing zero is not a test of equivalence.

Table 10. Held-out pre-projection benchmark; SCOUT includes affine calibration, and GRU uses the original 36-epoch budget. Each method uses identical test samples. RMSE is in source weight units; MAPE uses positive targets only. Bold values are the best within each cluster and metric.

Method	C1 Intermittent		C2 Regular		C3 Backbone	
	RMSE	MAPE (%)	RMSE	MAPE (%)	RMSE	MAPE (%)
Last value	2238.93	70.54	5948.59	53.51	16,918.11	31.12
Seasonal naive	2315.93	73.01	6840.95	59.81	19,203.47	35.61
Recent-4	1849.58	63.64	5217.19	46.66	14,938.40	28.09
Ridge direct	1806.01	65.60	5157.20	47.09	15,076.33	29.55
GBDT quantile	1811.30	64.83	5285.32	47.07	14,583.00	30.62
GRU direct (36 epochs)	1798.54	65.11	5124.55	46.92	14,406.85	28.60
SCOUT	1733.89	61.61	5047.18	43.47	14,505.98	28.39

5.7. Robustness, Stability, and Ablation Evidence

The expanded modern comparison in Table 11 compares pre-projection outputs. Later robustness and ablation tables explicitly identify their output stages; Appendix B documents the correction and diagnostic scope alongside the results reported in this section. SCOUT has lower RMSE than the Patch Transformer, two-stage intermittent model, and OD-graph diffusion baseline in C1 and C2. In C3, the Patch Transformer and GRU have lower RMSE, while the graph and two-stage methods do not. The result therefore does not establish superiority over all modern architectures.

Table 11. Modern and targeted pre-projection RMSE under the official split (seed 42 where stochastic). SCOUT includes affine calibration; GRU uses the additional-comparison 48-epoch budget.

Method	C1 Intermittent	C2 Regular	C3 Backbone
Recent-4	1849.58	5217.19	14,938.40
Two-stage intermittent	1935.90	5756.13	15,643.96
OD-graph diffusion	1777.34	5191.95	14,724.70
Patch Transformer	1827.62	5133.13	14,412.13
GRU direct (48 epochs)	1801.04	5150.21	14,389.36
SCOUT (calibrated)	1733.89	5047.18	14,505.98

Bold values indicate the lowest RMSE within each cluster.

Five independent training seeds show small SCOUT variation relative to GRU and Patch Transformer (Table 12). Across five chronological rolling-origin folds, the calibrated and projected SCOUT output has the lowest mean RMSE in all clusters, but these folds come from the same railway system and do not constitute external validation. Fold dates, sample counts, and individual RMSE values are detailed in Appendix B.

Table 12. Training-seed stability on the official split and mean RMSE over five rolling-origin folds. Seed entries are mean $\pm$ sample standard deviation over seeds 11, 23, 42, 67, and 101. SCOUT includes validation-based affine calibration and lower-bound projection, whereas comparator forecasts are reported before projection. These comparisons evaluate the reported pipelines and do not isolate forecasting-architecture effects.

Protocol	Method	C1	C2	C3
Five seeds	SCOUT	1731.74 $\pm$ 0.49	5050.60 $\pm$ 2.25	14,506.33 $\pm$ 1.71
Five seeds	GRU direct	1802.99 $\pm$ 2.32	5147.12 $\pm$ 11.79	14,436.43 $\pm$ 69.27
Five seeds	Patch Transformer	1818.21 $\pm$ 8.66	5113.51 $\pm$ 12.46	14,450.14 $\pm$ 23.75
Five rolling folds	Recent-4	1600.52	5338.40	14,809.15
Five rolling folds	GRU direct	1559.89	5247.71	14,430.70
Five rolling folds	SCOUT	1518.07	5056.36	14,234.41

Bold values indicate the lowest mean RMSE within each cluster in the five-rolling-fold comparison.

The complete pipeline was also refitted after shifting the train/validation/test boundaries by 56 and 28 days (Table 13). The calibrated and projected SCOUT output remains below Recent-4 in every cluster and window, with reductions from 2.56% to 6.91%. Its relation to GRU is not uniform: SCOUT is lower in seven of nine displayed cluster-window comparisons. In C3 at W-28 and W0, SCOUT exceeds GRU by 0.77% and 0.81%, respectively, using GRU as the denominator. Because the shifted windows overlap, this is a boundary-sensitivity test rather than three independent experiments.

Table 13. RMSE under shifted chronological boundaries. Profiles and models are refitted; GRU uses a 48-epoch budget. SCOUT includes validation-based affine calibration and lower-bound projection, whereas comparator forecasts are reported before projection. These comparisons evaluate the reported pipelines and do not isolate forecasting-architecture effects.

Window	Cluster	Recent-4	GRU Direct	SCOUT
W-56	C1	1728.16	1684.71	1608.83
W-56	C2	5227.95	5126.90	4979.82
W-56	C3	15,078.45	14,476.36	14,424.99
W-28	C1	1792.36	1742.02	1683.88
W-28	C2	5089.08	5027.59	4959.02
W-28	C3	14,701.73	14,202.22	14,311.24
W0	C1	1849.58	1801.04	1732.25
W0	C2	5217.19	5150.21	5049.90
W0	C3	14,938.40	14,389.36	14,506.00

Bold values indicate the lowest RMSE within each window and cluster.

One-factor threshold perturbations produce different route counts, so two views were retained. On each configuration's local route sets, SCOUT improves over its local Recent-4 reference in all 21 cluster-configuration cells: 5.59–6.82% in C1, 2.91–3.91% in C2, and 1.56–3.81% in C3. For Base-common routes, the reported relative change is defined as $100(\text{RMSE}_{\text{Base}} - \text{RMSE}_{\text{Variant}}) / \text{RMSE}_{\text{Base}}$; a positive value denotes lower RMSE for the variant. On Base-common routes with Base labels fixed, relative SCOUT RMSE reductions remain within -0.18 to $+0.07\%$ in C1, -0.04 to $+0.27\%$ in C2, and -0.78 to $+0.78\%$ in C3. Thus, the qualitative Recent-4 comparison is locally robust, but the default thresholds are not uniquely optimal.

Table 14 reports single-seed diagnostic comparisons under the original split. Each variant is fitted and calibrated under its stated specification and evaluated with the same standard lower-bound rule; the scenario map is the identity. After fixing the validation reference independently of the candidate bank, removal of the recent-level family yields RMSE values of 1772.27, 5138.76, and 14,874.62 in C1–C3, respectively, all higher than the full-model values. The global-weight and pooled-fit results remain competitive with the full model in some strata, and seasonal and trend/momentum removals have cluster-dependent effects. The two proxy-free variants both zero the eight proxy inputs: the 15-component variant replaces External with its Recent-14 reference, while the 14-component variant removes that candidate. Their difference concerns candidate representation and dimensionality under proxy-free inputs, not the isolated effect of removing External while retaining the proxy inputs. These diagnostics do not establish an independent accuracy gain from every module or identify unique temporal mechanisms.

Table 14. Diagnostic RMSE on the official split (seed 42). Variants are refitted, recalibrated, and evaluated after the same standard lower-bound rule, with no scenario adjustment. The validation reference is Recent-4 independently of candidate deletion. Both proxy-free variants zero eight proxy inputs; $K = 15$ replaces External with Recent-14 and $K = 14$ removes that candidate. The learned global mixture has no context adaptation.

Variant	C1	C2	C3
Full SCOUT + projection	1732.25	5049.90	14,506.00
Learned global simplex (no gate)	1732.07	5054.54	14,470.85
Proxy-free SCOUT ($K = 15$)	1731.99	5046.74	14,497.23
Unstratified pooled SCOUT	1732.12	5044.55	14,387.58
Without recent-level family	1772.27	5138.76	14,874.62
Without seasonal family	1748.33	5019.47	14,405.92
Without trend/momentum family	1731.84	5097.76	14,845.86
Proxy-free, External removed ($K = 14$)	1735.55	5044.63	14,471.74

5.8. Bootstrap Evidence and Multi-Horizon Behavior

All three RMSE bootstrap intervals exclude zero, although the C3 lower endpoint is close to zero (Table 15). The fractions of replicates with positive RMSE improvement are 1.000, 0.992, and 0.984 for C1–C3; these are resampling frequencies, not posterior probabilities. WAPE uncertainty includes zero for C1 and C2 but excludes zero for C3. MAPE uncertainty supports improvement in C1 and C2 but includes zero in C3. These differences reinforce the need to distinguish an RMSE result from a blanket performance claim. The date-block resampling preserves all routes and four horizons attached to each sampled forecast origin, thereby retaining within-date cross-route and cross-horizon dependence; it does not model dependence between separate route histories beyond their shared sampled dates.

Table 15. Paired seven-day moving-block bootstrap for SCOUT + projection relative to Recent-4 (2000 replicates). Positive WAPE and MAPE changes indicate improvement. The positive-improvement fraction is the proportion of bootstrap replicates with a positive RMSE reduction, not a posterior probability.

Cluster	RMSE Reduction (%)	95% Low	95% High	$P(\Delta_{RMSE} > 0)$	WAPE Change (pp)	MAPE Change (pp)
1	6.34	5.06	7.55	1.000	+0.00	+2.03
2	3.21	0.59	6.02	0.992	−0.13	+3.18
3	2.89	0.22	6.07	0.984	+0.55	−0.30

Alternative block lengths of 4, 7, and 14 days leave the RMSE conclusion unchanged. Across these choices, the 95% lower bounds for RMSE reduction are 5.06–5.12% in C1, 0.59–0.82% in C2, and 0.18–1.11% in C3; all nine intervals exclude zero. The seven-day value is retained because it spans a complete weekly operating cycle, not because it maximizes significance. Appendix B lists all nine intervals. All entries in those block-length intervals, Table 15, Figure 4, and the horizon intervals in Appendix B now use one canonical reanalysis (seed 20260923; clusters 1–3 and block lengths 4, 7, 14 processed in that order). Each replicate samples contiguous forecast-origin blocks with replacement and retains every route and horizon within each sampled date; the final concatenated draw is truncated to 171 origins. This preserves contemporaneous dependence and within-block temporal dependence, not dependence across artificial block joins or uncertainty from refitting the model. The intervals are percentile intervals conditional on the fitted predictions.

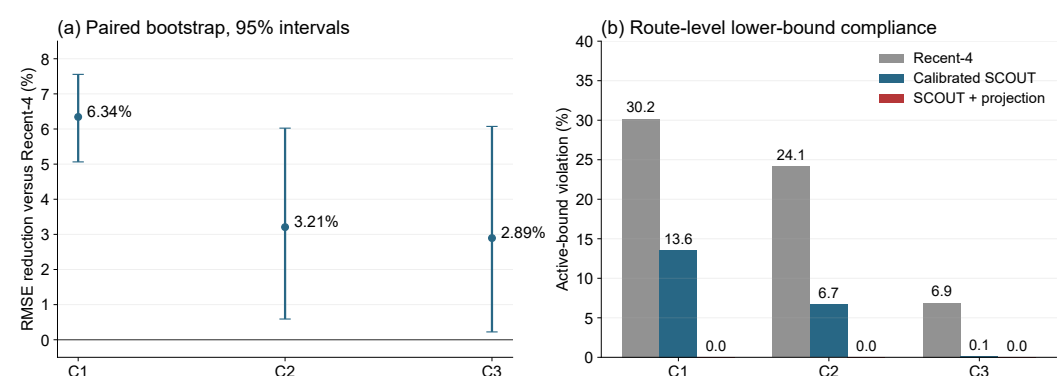


Figure 4. Accuracy and lower-bound-compliance evidence. Panel (a) gives paired seven-day moving-block bootstrap estimates and 95% intervals from 2000 replicates for the projected SCOUT forecast's RMSE reduction relative to Recent-4. Panel (b) reports policy violations conditional on a positive lower bound. The red SCOUT + projection series equals zero in every cluster; its bars therefore have zero height.

The lower-bound-compliant SCOUT forecast reduces RMSE at every horizon in all three clusters (Figure 5). The reduction ranges are 6.26–6.50% in C1, 2.29–4.32% in C2, and 1.56–3.50% in C3. MAPE improves at every horizon in C1 and C2, whereas C3 is modestly worse by at most 0.55 percentage points. The RMSE benefit persists through H4, but these results alone do not attribute that pattern to the horizon weights. The C3 relative-error trade-off remains. The pooled metrics use 21,546, 12,141, and 2223 route–origin cases (86,184, 48,564, and 8892 route–origin–horizon cells) in C1–C3, respectively; positive-target MAPE uses 8159, 26,653, and 8649 cells with $y > 0.1$.

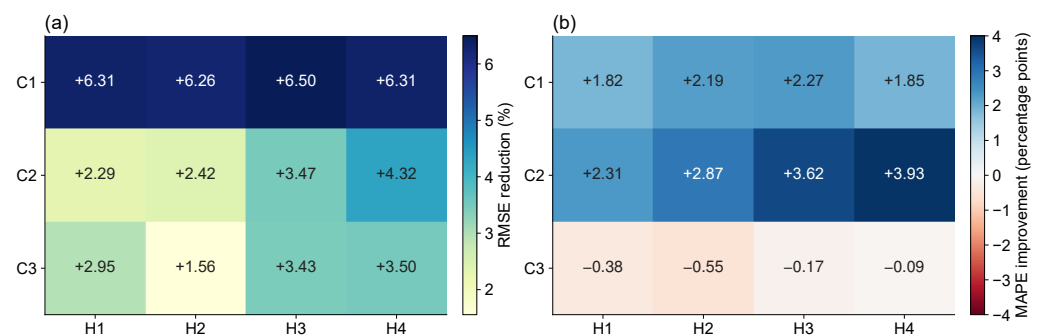


Figure 5. Horizon-specific changes of the lower-bound-compliant SCOUT forecast relative to Recent-4. Panel (a) reports RMSE reduction in percent; panel (b) reports MAPE improvement in percentage points. Positive entries indicate improvement. Differences use unrounded predictions, so subtraction of rounded table entries can differ by 0.01 percentage points.

5.9. Learned Component Contributions

Figure 6 reports contributions rather than raw gate weights. For each positive-component training context, a component’s weighted forecast is divided by the sum of all weighted components and then averaged over horizons. Positive-component contexts cover 46.28% of C1 training cases, 95.94% of C2, and 100.00% of C3. In C1, Recent-28 is largest at 24.71%, consistent with longer-window stabilization under intermittent activity. C2 is more distributed, with Last-Year largest at 8.05%. C3 places the largest share on Last Value (8.95%), followed by recent-level and momentum terms. These shares come from the actual horizon-specific prediction states; they do not imply a uniquely identified physical decomposition.

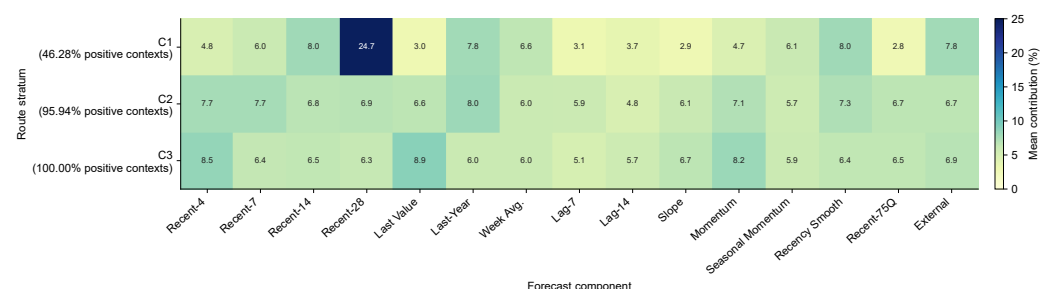


Figure 6. Mean weighted-output contribution shares, not raw simplex weights, for all 15 components, evaluated with horizon-specific validation-selected states. Row labels report the fraction of training contexts with a positive total component forecast. Shares are conditional on those contexts and sum to 100% within each cluster.

The contribution ordering is numerically stable across the five seeds: Spearman rank correlations against seed 42 are generally near 0.99 and the observed minima remain between 0.95 and 0.97 across clusters and horizons. Stability does not imply uniqueness, however. Same-route, same-horizon Spearman diagnostics show median absolute pair correlations of 0.38–0.42 in C1, 0.54–0.56 in C2, and 0.57–0.60 in C3, with maxima of 0.98 or

higher. Correlated candidates can substitute for one another, so Figure 6 is an auditable record of model use, not a causal decomposition or evidence that human users interpret SCOUT better than another model.

5.10. Representative Route Behavior

One route per cluster was selected before examining test errors: the route closest to its cluster medians in standardized train-period log mean and non-zero ratio. This rule avoids choosing visually favorable examples. Figure 7 displays observed test flows, H1 and H4 SCOUT + projection forecasts, and the H4 validation-residual band. C1 remains difficult because the model often predicts a smoothed positive level while actual activations are sparse and spiky. C2 tracks the operating level more consistently but does not reproduce every abrupt change. C3 follows the broad level of a continuously active route; its large scale makes absolute error relevant, but realized operational benefit has not been measured.

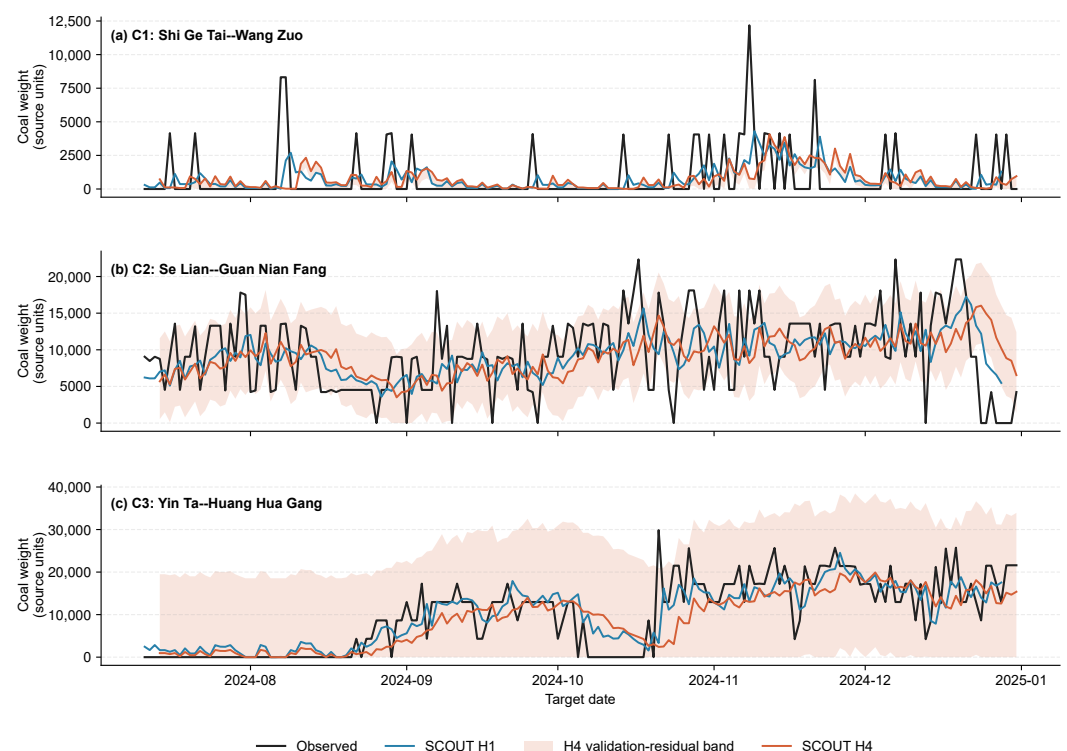


Figure 7. Held-out forecasts for routes selected from training-period profile medians independently of test error: C1: Shi Ge Tai–Wang Zuo; C2: Se Lian–Guan Nian Fang; C3: Yin Ta–Huang Hua Gang. Vertical scales are cluster-specific and use the source weight unit. The H4 band uses pooled validation residual quantiles and is descriptive rather than a conditional coverage guarantee.

5.11. Scenario Stress Test and Interval Diagnostics

The market-shock setting lowers the selected backbone mean forecast by 6.50%. The supply-security setting raises it by 10.69%. In the combined case, the scenario lower bound dominates the downward market multiplier, producing the same lower-bound-compliant mean as the supply-security case. This is the intended operator precedence, but the magnitude is illustrative because the input indices and response schedule are designed rather than estimated from observed markets.

Table 16 summarizes the scenario outcomes, and Figure 8 shows the corresponding four-horizon paths.

Table 16. Illustrative scenario stress test for the highest-training-mean backbone route. Mean forecasts are in source weight units. S0 is the calibrated forecast before either standard-policy or scenario projection.

Scenario	Mean Forecast	Change from S0 (%)
S0 Baseline	110,631.35	+0.00
S1 Market Shock	103,439.92	−6.50
S2 Supply Security	122,455.23	+10.69
S3 Combined	122,455.23	+10.69

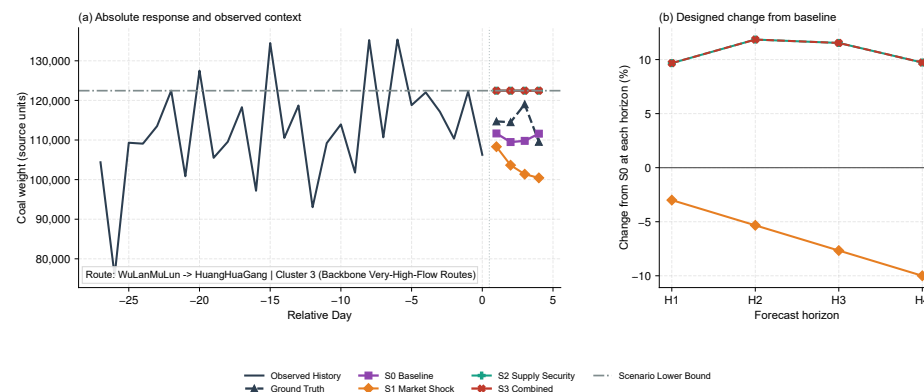


Figure 8. Illustrative four-horizon response for the highest-training-mean backbone route, Wu Lan Mu Lun–Huang Hua Gang, at forecast origin 31 October 2024. Panel (a) shows the 28-day observed context, held-out truth, and forecast paths in source weight units; panel (b) shows the designed horizon-specific change from S0. S0 is the calibrated forecast before projection; the scenario multiplier and lower bound are documented stress-test inputs rather than causal market estimates.

The nominal residual band is marginally close to 80% overall in every cluster (Table 17), but this aggregate result is misleading for C1: zero targets dominate, and coverage conditional on positive targets is only 4.73%. C2 positive-target coverage is 74.14%, while C3 reaches 80.09%. The C1 interval should therefore not be used as a reliable uncertainty estimate for rare activations. A two-stage activation/volume model and suitable conditional calibration are directions for improvement, not guarantees established here. Adaptive conformal inference offers a relevant approach to changing distributions [34], but its long-run coverage goal does not by itself guarantee positive-event coverage for C1.

Table 17. Validation-residual interval diagnostics on the held-out test. The nominal marginal level is 80%.

Cluster	Positive Targets (%)	Overall Coverage (%)	Positive-Target Coverage (%)	Mean Width	Width/Positive Mean (%)
1	9.47	80.36	4.73	188.28	2.91
2	54.88	81.72	74.14	8646.92	73.47
3	97.27	80.44	80.09	33,509.86	61.63

6. Discussion

6.1. What the Evidence Supports

The strongest empirical finding is lower RMSE than Recent-4 in the reported pipeline comparisons. In Tables 12 and 13 and the rolling-fold details in Appendix B, these comparisons do not separate forecasting-architecture effects from calibration and lower-bound projection. Pooled RMSE-reduction intervals exclude zero in all three clusters for block lengths of 4, 7, and 14 days. All horizon-specific point estimates of RMSE reduction are positive, but the corresponding 95% intervals for C2-H2 and C3-H2 include zero, so not every

horizon-specific difference is statistically distinguished. The same directional comparison holds in five rolling folds, two shifted temporal boundaries, and all one-factor threshold configurations. The modern benchmark narrows that claim: SCOUT is strongest in C1–C2, whereas GRU or Patch Transformer can have lower C3 RMSE. This can matter for railway planning because high-volume-route deviations enter train and terminal decisions. The resulting effects on operational decision quality are not directly evaluated here, and the confidential source unit prevents conversion of forecast differences into wagons, trainloads, cost, or capacity deviations. These RMSE gains do not extend to every metric: C2 WAPE is 0.13 percentage points worse than Recent-4, C3 MAPE is 0.30 points worse, and sMAPE increases in all three clusters (Appendix A).

The cluster design clarifies this trade-off. C1 contains enough activity for a continuous volume model but remains highly intermittent; its WAPE above 100% and poor positive-event interval coverage show that the present formulation has not solved activation timing. C2 benefits most consistently in positive-target MAPE and later-horizon RMSE, but its total absolute error is essentially tied with persistence. C3 is almost continuously active, so relative metrics are stable and absolute-error reduction is the principal gain. These distinct regimes should not be compressed into one network-wide percentage.

6.2. Structural Interpretability and Decision Transparency

SCOUT's structural interpretability comes from explicitly defined reference components representing overlapping longer-term tendencies, periodic variation, and short-term changes. The four functional groups organize their implementation and diagnostics, while context gating adapts their combination. Contribution and stability analyses describe how the fitted predictor uses this representation; they neither create its temporal meaning nor demonstrate uniquely identified causal mechanisms or superior user comprehension.

A separate form of transparency follows from retaining the distinction between the statistical forecast, declared scenario adjustments, and route-level lower-bound enforcement. Reporting Recent-4 and SCOUT before and after the same standard bounds makes changes in point error and rule compliance visible. Several simpler variants achieve similar or lower RMSE than full SCOUT in Table 14. These refitted and recalibrated single-seed comparisons do not establish an accuracy benefit from every module in every stratum. SCOUT's contribution lies in its explicit temporal representation and the structured integration of adaptive combination, scenario adjustment, and auditable lower-bound enforcement. The projected output is compliant with the stated lower bounds only, not a demonstrated network-feasible operating plan.

6.3. Limitations and Generalizability

The findings are subject to several limitations. First, the study evaluates one railway coal OD system. Rolling folds and shifted boundaries test temporal sensitivity within that system but do not establish cross-network or future-regime validity. Second, the known-input indices are generated proxies. Their formulas are reproducible, but they cannot support economic attribution, causal interpretation, or claims about actual coal-price elasticity. Third, the route-wise projection enforces only lower bounds and can remain infeasible at network level when shared corridor, loading, unloading, fleet, or conservation constraints are considered. Fourth, residual intervals are pooled marginal bands and perform poorly for positive C1 events. Fifth, route categories are useful for regime-specific reporting, but the pooled-fit result shows that stratification is not intrinsically accuracy-improving; deployment should monitor distribution drift and re-estimate eligibility and strata without using future outcomes.

The structural interpretability described in Section 6.2 has not been assessed through comparative experiments or user studies. Evaluating user comprehension relative to alternative neural forecasting models remains a direction for future research.

Independent reproducibility is a further limitation: the underlying railway data and implementation code cannot be shared. The manuscript specifies the method and reports aggregate diagnostics, but it does not provide an executable replication or permit independent verification of the empirical results from the manuscript alone.

Future work should systematically compare alternative component banks and combine an activation classifier with a conditional volume model for intermittent routes. Other priorities include replacing proxy indices with timestamped observed covariates with documented availability checks, adding graph or bipartite route dependencies, calibrating distribution-free intervals by cluster and horizon, and projecting forecasts onto a network-level feasible set. These extensions should be judged with the same separation of predictive accuracy, scenario response, and operational feasibility.

7. Conclusions

This study developed SCOUT as a structurally organized forecasting approach for eligible railway coal OD flows. Explicit reference components implement overlapping objectives for longer-term tendencies, periodic variation, and short-term changes, while adaptive gating and separate scenario and lower-bound adjustments perform distinct computational and decision-processing roles. Across 210 train-period-eligible routes representing 98.42% of held-out network flow, the lower-bound-compliant SCOUT forecast reduced RMSE relative to Recent-4 by 6.34%, 3.21%, and 2.89%. The direction persists across rolling folds, shifted boundaries, threshold perturbations, and alternative bootstrap block lengths. The route-level projection eliminated all active-bound violations, while its effect on point accuracy was mixed and is reported separately.

Against conventional, intermittent, graph, recurrent, and patch-transformer references, SCOUT is most competitive in C1–C2. In C3, GRU and Patch Transformer can achieve lower RMSE. Five-seed variation is small for SCOUT, yet gate, proxy, seasonal-family, and stratification ablations have mixed directions. The evidence therefore supports cluster-dependent competitiveness and stronger large-error control, not universal superiority or independent benefit from every module.

Positive-target MAPE improves in C1 and C2 but not C3; C2 WAPE is essentially unchanged; sMAPE worsens; and the C1 marginal interval fails to cover most positive events. The supported conclusion is that SCOUT improves large-error control and provides an explicit, auditable interface for scenario and lower-bound decisions on eligible routes in this dataset. Reliable network-wide activation, causal external drivers, calibrated conditional uncertainty, and network-capacity feasibility remain open requirements for operational deployment.

Author Contributions: Conceptualization, J.Z.; methodology, J.Z. and J.C.; software, J.Z.; validation, J.Z., J.C. and Z.O.; formal analysis, J.Z. and Z.O.; data curation, J.Z.; writing—original draft preparation, J.Z.; writing—review and editing, J.C. and Z.O.; visualization, J.Z.; supervision, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The underlying railway data are subject to the data provider's access restrictions and cannot be shared. The authors are also unable to share the implementation code.

Method definitions, evaluation protocols, parameter settings, and aggregate results are reported within the manuscript, but independent reproduction of the empirical results is not possible from these materials alone.

Acknowledgments: The authors thank the data provider for supporting this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OD	Origin–destination
SCOUT	Scenario-Oriented Context-Gated Forecasting
MAE	Mean absolute error
RMSE	Root mean square error
MAPE	Mean absolute percentage error
WAPE	Weighted absolute percentage error
sMAPE	Symmetric mean absolute percentage error

Appendix A. Additional Metrics and Definitions

Symmetric MAPE is included as a robustness diagnostic because zero and near-zero cells can make percentage metrics behave differently. The overall violation column uses all test cells and is shown only to illustrate why the conditional denominator in the main text is preferable. Pinball losses are reported only for SCOUT + projection because the deterministic baselines do not define quantiles.

Table A1 reports the additional diagnostics; Table A2 gives the horizon-specific point metrics.

Table A1. Additional test diagnostics for Recent-4 and SCOUT + projection.

Cluster	Variant	sMAPE (%)	All-Cell Violation (%)	Pinball $q_{0.1}$	Pinball $q_{0.9}$
1	Recent-4	156.05	0.160	–	–
1	SCOUT + projection	182.95	0.000	186.31	314.82
2	Recent-4	92.84	0.230	–	–
2	SCOUT + projection	114.69	0.000	649.42	982.02
3	Recent-4	25.48	0.540	–	–
3	SCOUT + projection	27.98	0.000	2612.51	2676.03

Table A2. Horizon-specific held-out RMSE and positive-target MAPE. SCOUT denotes the lower-bound-projected output.

Cluster	Horizon	Recent-4 RMSE	SCOUT RMSE	Recent-4 MAPE (%)	SCOUT MAPE (%)
1	H1	1790.95	1678.01	61.64	59.82
1	H2	1839.54	1724.43	63.52	61.33
1	H3	1877.87	1755.87	64.54	62.27
1	H4	1888.39	1769.26	64.86	63.02
2	H1	4800.13	4690.16	44.01	41.71
2	H2	5106.39	4982.95	46.26	43.38
2	H3	5351.46	5165.76	47.52	43.90
2	H4	5578.69	5337.91	48.88	44.95
3	H1	13,709.94	13,305.59	25.79	26.17
3	H2	14,662.69	14,434.53	27.66	28.21
3	H3	15,388.07	14,859.69	28.91	29.08
3	H4	15,901.99	15,345.72	29.99	30.08

For test cells Ω and positive cells $\Omega^+ = \{j : y_j > 0.1\}$, the metrics are defined in Equations (A1)–(A4):

$$\text{MAE} = |\Omega|^{-1} \sum_{j \in \Omega} |y_j - \hat{y}_j|, \quad (\text{A1})$$

$$\text{RMSE} = \left(|\Omega|^{-1} \sum_{j \in \Omega} (y_j - \hat{y}_j)^2 \right)^{1/2}, \quad (\text{A2})$$

$$\text{MAPE} = 100 |\Omega^+|^{-1} \sum_{j \in \Omega^+} \left| \frac{y_j - \hat{y}_j}{y_j} \right|, \quad (\text{A3})$$

$$\text{WAPE} = 100 \frac{\sum_{j \in \Omega} |y_j - \hat{y}_j|}{\sum_{j \in \Omega} |y_j|}. \quad (\text{A4})$$

The fixed cutoff of 0.1 excludes zero and very small denominators from positive-target MAPE. It is a numerical evaluation convention, not a physical operating threshold. At the route level, MAPE is undefined when no positive test target exists and is therefore not replaced by zero. WAPE can exceed 100% in C1 because the sum of absolute misses, including positive forecasts on zero-demand cells, can exceed the small sum of observed intermittent demand; this makes it a descriptive error ratio rather than a stable service-level measure for that stratum. The forecast origin is the last observed day, the forecast-window start is the first target day, and the target date is defined separately for each horizon.

Appendix B. Additional Diagnostics and Implementation Clarifications

The recent-level family ablation in Table 14 required a validation-reference correction. Removing that family changed candidate index zero, which had also supplied the normalization reference for validation scoring. The corrected diagnostic uses independently stored Recent-4 forecasts for scoring, without reintroducing Recent-4 into the ablated candidate bank. Only the affected C1–C3 fits were repeated under the same split, seed 42, input settings, and selection policy. The resulting RMSE values are 1772.27, 5138.76, and 14,874.62, respectively. The selected H1 epoch changed from 9 to 10 in C1 and from 16 to 17 in C2; C3 selections and predictions were unchanged. Checks confirmed matching configurations, evaluation sample keys, observed targets, baseline predictions, and lower bounds. Other family removals retain Recent-4 at candidate index zero; their reference values were therefore unaffected, and those fits and the full-model experiments were not rerun for this correction.

For local loss-weight sensitivity, the default weights on the relative-error, log-Huber, and squared-error terms are (0.58, 0.32, 0.10). Five alternative triples were evaluated with the official split and seed 42: (0.333333, 0.333333, 0.333334), (0.25, 0.65, 0.10), (0.85, 0, 0.15), (0.75, 0.15, 0.10), and (0.30, 0.20, 0.50). In that order, the lower-bound-compliant RMSE vectors for C1–C3 are (1732.25, 5049.58, 14,505.56), (1732.32, 5049.58, 14,505.43), (1732.13, 5049.90, 14,506.13), (1732.17, 5049.90, 14,506.10), and (1732.25, 5049.57, 14,505.37). Relative to the default (1732.25, 5049.90, 14,506.00), the absolute relative change is below 0.01% in every reported cluster-setting comparison. These are local sensitivity diagnostics, not a test-set search for a replacement default or evidence that these weights are universally optimal.

The proxy formulas, model-specific protocols, sensitivity settings, and aggregate results are documented within the manuscript. The underlying railway data and implementation code are not available for sharing. This documentation supports methodological inspection but does not constitute an independently executable replication. The reproducibility limitation is also stated in Section 6.3 and the Data Availability Statement.

The fixed-state calibration diagnostic in Table A3 disables only post-fit affine calibration and leaves the selected neural mixture unchanged. Calibration lowers RMSE in all three clusters, but not every other metric (C1 MAPE increases slightly). This is distinct from retraining a model under a different objective. The standard-policy forecast uses the identity scenario map; the illustrative non-identity scenario remains separately reported in Table 16.

Table A3. Stage-specific metrics on the identical official test cells (seed 42).

Cluster	Output Stage	RMSE	MAE	MAPE (%)
1	Uncalibrated mixture	1735.44	626.70	61.56
1	Affine-calibrated	1733.89	623.45	61.61
1	Calibrated + lower bound	1732.25	623.36	61.61
2	Uncalibrated mixture	5059.28	3136.41	44.04
2	Affine-calibrated	5047.18	3131.03	43.47
2	Calibrated + lower bound	5049.90	3132.20	43.48
3	Uncalibrated mixture	14,562.03	10,556.06	28.94
3	Affine-calibrated	14,505.98	10,484.30	28.39
3	Calibrated + lower bound	14,506.00	10,484.31	28.39

Table A4 lists the paired intervals for all three block lengths.

Table A4. Block-length sensitivity of paired RMSE reduction relative to Recent-4.

Cluster	Block Days	Reduction (%)	95% Low	95% High
1	4	6.34	5.06	7.52
1	7	6.34	5.06	7.55
1	14	6.34	5.12	7.51
2	4	3.21	0.76	5.69
2	7	3.21	0.59	6.02
2	14	3.21	0.82	6.18
3	4	2.89	0.18	5.86
3	7	2.89	0.22	6.07
3	14	2.89	1.11	6.08

Section 5.8 reports the total and positive-target cell counts; Section 5.5 and Table 9 report the policy-active denominators and projection diagnostics. Tables A4 and A5 provide the block-length and horizon-specific comparative intervals reported here. Confidence intervals for differences in forecasting error do not establish predictive coverage for individual positive-flow events.

Table A5. Horizon-specific paired seven-day bootstrap intervals for projected SCOUT versus Recent-4. RMSE reductions are percent; MAPE changes are percentage points. Positive values favor SCOUT.

Cluster	Horizon	RMSE Reduction	95% Interval	MAPE Change	95% Interval
1	H1	6.31	[4.79, 7.72]	+1.82	[0.72, 2.86]
1	H2	6.26	[4.51, 7.84]	+2.19	[1.11, 3.47]
1	H3	6.50	[4.81, 8.09]	+2.27	[1.16, 3.60]
1	H4	6.31	[4.48, 8.07]	+1.85	[0.53, 3.21]
2	H1	2.29	[0.06, 4.57]	+2.31	[1.79, 2.91]
2	H2	2.42	[−0.24, 5.17]	+2.87	[2.25, 3.62]
2	H3	3.47	[0.43, 6.84]	+3.62	[2.90, 4.50]
2	H4	4.32	[1.55, 7.58]	+3.93	[3.25, 4.76]
3	H1	2.95	[0.57, 5.79]	−0.38	[−0.89, 0.69]
3	H2	1.56	[−1.33, 5.12]	−0.55	[−1.30, 0.93]
3	H3	3.43	[0.62, 6.80]	−0.17	[−0.96, 1.22]
3	H4	3.50	[0.39, 6.88]	−0.09	[−1.17, 1.32]

Tables A6 and A7 detail the five expanding-training rolling-origin folds summarized in Table 12. All training/profile periods start on 1 November 2021. Each validation and test block contains 90 target dates. Retaining only samples whose four target dates fall within the assigned block yields 87 validation and 87 test forecast origins per route. Route eligibility and strata are recomputed from each fold’s training observations; model fitting, selection, and calibration use training and validation data only. These folds are distinct from the three shifted-boundary windows in Table 13.

Table A6. Rolling-fold target-date boundaries (inclusive). Training periods share the start date 1 November 2021.

Fold	Training End	Validation Dates	Test Dates
R1	10 July 2023	11 July 2023 to 8 October 2023	9 October 2023 to 6 January 2024
R2	8 October 2023	9 October 2023 to 6 January 2024	7 January 2024 to 5 April 2024
R3	6 January 2024	7 January 2024 to 5 April 2024	6 April 2024 to 4 July 2024
R4	5 April 2024	6 April 2024 to 4 July 2024	5 July 2024 to 2 October 2024
R5	4 July 2024	5 July 2024 to 2 October 2024	3 October 2024 to 31 December 2024

Table A7. Rolling-fold test samples and RMSE (seed 42). Cases are route–origin pairs, each with four horizon observations. SCOUT includes validation-based affine calibration and lower-bound projection, whereas comparator forecasts are reported before projection. These comparisons evaluate the reported pipelines and do not isolate forecasting-architecture effects. The output stages are consistent with Table 12.

Fold	Cluster	Routes	Test Cases	Recent-4	GRU	SCOUT
R1	1	121	10,527	1432.29	1386.81	1383.18
R1	2	70	6090	5471.68	5424.73	5293.73
R1	3	13	1131	13,462.20	13,405.91	13,136.46
R2	1	123	10,701	1653.65	1631.22	1573.76
R2	2	69	6003	5766.45	5662.18	5400.93
R2	3	13	1131	15,040.96	14,451.37	14,402.03
R3	1	126	10,962	1664.09	1621.87	1564.04
R3	2	70	6090	5540.06	5391.13	5115.94
R3	3	13	1131	15,545.47	14,798.72	14,631.13
R4	1	126	10,962	1487.46	1454.78	1421.17
R4	2	71	6177	4678.59	4636.42	4478.87
R4	3	13	1131	14,900.81	14,842.92	14,357.02
R5	1	126	10,962	1765.13	1704.75	1648.19
R5	2	71	6177	5235.23	5124.10	4992.32
R5	3	13	1131	15,096.30	14,654.58	14,645.42

The rolling-fold RMSE means in Table 12 are unweighted arithmetic means of the five fold-specific RMSE values, not RMSE computed after pooling all test cells. The 15 fold-by-cluster comparisons concern these three methods in one railway system and are not evidence of universal superiority or external generalization.

References

1. International Energy Agency. *Coal 2025: Analysis and Forecast to 2030*; Technical Report; International Energy Agency: Paris, France, 2025.

2. Crainic, T.G. Service network design in freight transportation. *Eur. J. Oper. Res.* **2000**, *122*, 272–288. [https://doi.org/10.1016/S0377-2217\(99\)00233-7](https://doi.org/10.1016/S0377-2217(99)00233-7).
3. Macharis, C.; Bontekoning, Y.M. Opportunities for OR in intermodal freight transport research: A review. *Eur. J. Oper. Res.* **2004**, *153*, 400–416. [https://doi.org/10.1016/S0377-2217\(03\)00161-9](https://doi.org/10.1016/S0377-2217(03)00161-9).
4. de Jong, G.; Ben-Akiva, M. A micro-simulation model of shipment size and transport chain choice. *Transp. Res. Part B Methodol.* **2007**, *41*, 950–965. <https://doi.org/10.1016/j.trb.2007.05.002>.
5. Gédéon, C.; Florian, M.; Crainic, T.G. Determining origin-destination matrices and optimal multiproduct flows for freight transportation over multimodal networks. *Transp. Res. Part B Methodol.* **1993**, *27*, 351–368. [https://doi.org/10.1016/0191-2615\(93\)90022-3](https://doi.org/10.1016/0191-2615(93)90022-3).
6. Hwang, T.; Ouyang, Y. Assignment of freight shipment demand in congested rail networks. *Transp. Res. Rec. J. Transp. Res. Board* **2014**, *2448*, 37–44. <https://doi.org/10.3141/2448-05>.
7. Wan, Y.; Li, W. The coal origin-destination matrix analysis and multimodal transportation cost modelling in the Yangtze River region, China. *Syst. Sci. Control Eng.* **2018**, *6*, 57–63. <https://doi.org/10.1080/21642583.2018.1518164>.
8. Liu, C.; Zhang, J.; Luo, X.; Yang, Y.; Hu, C. Railway Freight Demand Forecasting Based on Multiple Factors: Grey Relational Analysis and Deep Autoencoder Neural Networks. *Sustainability* **2023**, *15*, 9652. <https://doi.org/10.3390/su15129652>.
9. Cascetta, E. Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. *Transp. Res. Part B Methodol.* **1984**, *18*, 289–299. [https://doi.org/10.1016/0191-2615\(84\)90012-2](https://doi.org/10.1016/0191-2615(84)90012-2).
10. Hazelton, M.L. Inference for origin-destination matrices: Estimation, prediction and reconstruction. *Transp. Res. Part B Methodol.* **2001**, *35*, 667–676. [https://doi.org/10.1016/S0191-2615\(00\)00009-6](https://doi.org/10.1016/S0191-2615(00)00009-6).
11. Ashok, K.; Ben-Akiva, M.E. Estimation and Prediction of Time-Dependent Origin-Destination Flows with a Stochastic Mapping to Path Flows and Link Flows. *Transp. Sci.* **2002**, *36*, 184–198. <https://doi.org/10.1287/trsc.36.2.184.563>.
12. Zhou, X.; Mahmassani, H.S. A structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. *Transp. Res. Part B Methodol.* **2007**, *41*, 823–840. <https://doi.org/10.1016/j.trb.2007.02.004>.
13. Xiong, X.; Ozbay, K.; Jin, L.; Feng, C. Dynamic Origin-Destination Matrix Prediction with Line Graph Neural Networks and Kalman Filter. *Transp. Res. Rec. J. Transp. Res. Board* **2020**, *2674*, 491–503. <https://doi.org/10.1177/0361198120919399>.
14. Rong, C.; Ding, J.; Li, Y. An Interdisciplinary Survey on Origin-Destination Flows Modeling: Theory and Techniques. *Acm Comput. Surv.* **2025**, *57*, 1–49. <https://doi.org/10.1145/3682058>.
15. Zhang, J.; Che, H.; Chen, F.C.; Ma, W.; He, Z. Short-term origin-destination demand prediction in urban rail transit systems: A channel-wise attentive split-convolutional neural network method. *Transp. Res. Part C Emerg. Technol.* **2021**, *124*, 102928. <https://doi.org/10.1016/j.trc.2020.102928>.
16. Jiang, W.; Ma, Z.; Koutsopoulos, H.N. Deep learning for short-term origin-destination passenger flow prediction under partial observability in urban railway systems. *Neural Comput. Appl.* **2022**, *34*, 4813–4830. <https://doi.org/10.1007/s00521-021-06669-1>.
17. Vlahogianni, E.I.; Karlaftis, M.G.; Golias, J.C. Short-term traffic forecasting: Where we are and where we're going. *Transp. Res. Part C Emerg. Technol.* **2014**, *43*, 3–19. <https://doi.org/10.1016/j.trc.2014.01.005>.
18. Ermagun, A.; Levinson, D. Spatiotemporal traffic forecasting: Review and proposed directions. *Transp. Rev.* **2018**, *38*, 786–814. <https://doi.org/10.1080/01441647.2018.1442887>.
19. Tedjopurnomo, D.A.; Bao, Z.; Zheng, B.; Choudhury, F.M.; Qin, A.K. A Survey on Modern Deep Neural Network for Traffic Prediction: Trends, Methods and Challenges. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 1544–1561. <https://doi.org/10.1109/TKDE.2020.3001195>.
20. Li, Y.; Yu, R.; Shahabi, C.; Liu, Y. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
21. Yu, B.; Yin, H.; Zhu, Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, 13–19 July 2018; pp. 3634–3640. <https://doi.org/10.24963/ijcai.2018/505>.
22. Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Zhang, C. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Macao, China, 10–16 August 2019; pp. 1907–1913. <https://doi.org/10.24963/ijcai.2019/264>.
23. Zheng, C.; Fan, X.; Wang, C.; Qi, J. GMAN: A Graph Multi-Attention Network for Traffic Prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 1234–1241. <https://doi.org/10.1609/aaai.v34i01.5477>.
24. Bai, L.; Yao, L.; Li, C.; Wang, X.; Wang, C. Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 17804–17815.
25. Jiang, W.; Luo, J. Graph neural network for traffic forecasting: A survey. *Expert Syst. Appl.* **2022**, *207*, 117921. <https://doi.org/10.1016/j.eswa.2022.117921>.

26. Ni, R.; Lin, Z.; Wang, S.; Fanti, G. Mixture-of-Linear-Experts for Long-Term Time Series Forecasting. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 2–4 May 2024*; *Proceedings of Machine Learning Research*; Volume 238, pp. 4672–4680.
27. Balderas-Díaz, S.; Guerrero-Contreras, G.; Muñoz, A.; Boubeta-Puig, J. Fusing Temporal and Contextual Features for Enhanced Traffic Volume Prediction. In *Good Practices and New Perspectives in Information Systems and Technologies; Lecture Notes in Networks and Systems*; Springer: Cham, Switzerland, 2024; Volume 986, pp. 74–84. https://doi.org/10.1007/978-3-031-60218-4_8.
28. Yang, Y.; Ding, C.; Lee, S.; Yu, L.; Ma, F. A Modified Teunter–Syntetos–Babai Method for Intermittent Demand Forecasting. *J. Manag. Sci. Eng.* **2021**, *6*, 53–63. <https://doi.org/10.1016/j.jmse.2021.02.008>.
29. Kourentzes, N.; Athanasopoulos, G. Elucidate Structure in Intermittent Demand Series. *Eur. J. Oper. Res.* **2021**, *288*, 141–152. <https://doi.org/10.1016/j.ejor.2020.05.046>.
30. Hyndman, R.J.; Ahmed, R.A.; Athanasopoulos, G.; Shang, H.L. Optimal combination forecasts for hierarchical time series. *Comput. Stat. Data Anal.* **2011**, *55*, 2579–2589. <https://doi.org/10.1016/j.csda.2011.03.006>.
31. Wickramasuriya, S.L.; Athanasopoulos, G.; Hyndman, R.J. Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *J. Am. Stat. Assoc.* **2019**, *114*, 804–819. <https://doi.org/10.1080/01621459.2018.1448825>.
32. Wickramasuriya, S.L.; Turlach, B.A.; Hyndman, R.J. Optimal non-negative forecast reconciliation. *Stat. Comput.* **2020**, *30*, 1167–1182. <https://doi.org/10.1007/s11222-020-09930-0>.
33. Athanasopoulos, G.; Hyndman, R.J.; Kourentzes, N.; Panagiotelis, A. Forecast reconciliation: A review. *Int. J. Forecast.* **2024**, *40*, 430–456. <https://doi.org/10.1016/j.ijforecast.2023.10.010>.
34. Gibbs, I.; Candès, E. Adaptive Conformal Inference Under Distribution Shift. In *Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021*; Volume 34, pp. 1660–1672.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.