AIMS Mathematics, 11(10): 32628–32670.
DOI: 10.3934/math.20261282
Received: 01 August 2026
Revised: 14 September 2026
Accepted: 22 September 2026
Published: 10 October 2026

https://www.aimspress.com/journal/Math

Research article

A conformable fractional sparse robust grey RVFL model for energy time series forecasting

Yang Yue* **and Shaoyong Liu**

School of Big Data and Statistics, Tongling University, Tongling 244061, China

* **Correspondence:** Email: 2023003@tlu.edu.cn.

Abstract: Forecasting short energy time series is challenging because limited modeling observations, nonlinear dynamics, and atypical residuals can destabilize conventional grey models. This study proposed a conformable fractional sparse robust grey random vector functional link model (CFSRGRVFL). The model embeded a complete random vector functional link (RVFL) forcing term, comprising an unpenalized direct time link and randomized hidden features, within a conformable fractional grey system to control temporal weighting and nonlinear representation. Coefficients were estimated using a bounded exponential-squared loss, while an ℓ_1/ℓ_2 regularizer was applied only to the hidden-output coefficients to promote sparse feature selection without suppressing the direct link. An alternating direction method of multipliers scheme, incorporating an ℓ_1/ℓ_2 proximal update and difference-of-convex majorization, was developed for numerical estimation. Bayesian optimization with a fixed seed set tuned the coupled hyperparameters using the complete modeling sequence without accessing forecasting observations. Experiments on five energy series against three neural grey benchmarks showed that CFSRGRVFL achieved the lowest forecasting mean absolute percentage error in all cases and ranks first on most out-of-sample metrics. These results supported the model's effectiveness within the tested short-series settings.

Keywords: short time series forecasting; conformable fractional accumulation; random vector functional link; exponential-squared loss; ℓ_1/ℓ_2 regularization; bayesian optimization
Mathematics Subject Classification: 26A33, 62M10, 68T05

1. Introduction

Reliable energy forecasting often confronts two difficulties: the available series is short, and individual observations may be affected by structural change or measurement error. Grey system theory provides a low-dimensional framework for systems with limited and partially known information [1,2] and has been applied to energy consumption [3,4], renewable-energy generation [5],

non-renewable energy consumption [6], and photovoltaic forecasting [7]. However, when nonuniform temporal information, nonlinear variation, and potentially outlying observations occur together, the accumulation operator, nonlinear representation, and coefficient estimator used in conventional grey models still leave room for improvement.

Accumulated generation is a defining component of grey forecasting. Conventional models generally use first-order accumulated generation to attenuate fluctuations and expose the underlying trend. Extensions based on alternative discretizations and grey-equation optimization [8, 9], modified initial conditions [10], and deterministic forcing terms [11] improve different parts of the grey equation, but they retain a fixed cumulative weighting mechanism. Fractional accumulation introduces an order parameter that changes the contribution of historical observations [12–14]. In particular, conformable fractional accumulation (CFA) and CFD provide explicit order-dependent weights for transformation and restoration and have been incorporated into conformable fractional and nonhomogeneous grey models [15–17] and, more recently, into kernel grey models [18]. Existing conformable fractional models, however, mainly combine these operators with parametric or kernel forcing terms rather than a complete random vector functional link (RVFL) representation with both a direct link and selectable random hidden features.

Representing the unknown nonlinear forcing term is another central issue. Parametric nonlinear grey models retain a compact structure but require the forcing law to be specified in advance [19–21]. A related multivariable grey differential dynamic prediction model with new structures is presented in [22]. Kernel grey models relax this requirement through implicit feature mappings [23–25]; multi-kernel extensions are given in [26], while neural grey models learn nonlinear representations from the data [27–29]. RVFL offers a different balance between representation and estimation: it preserves a direct input–output link, uses randomized hidden features to represent nonlinear departures, and keeps the output stage linear in its coefficients [30, 31]. RVFL models have been applied to spatiotemporal processes and renewable-energy forecasting [32–34], and have recently been embedded in grey forecasting [35]. Existing grey RVFL formulations, however, primarily rely on ordinary accumulation and squared-loss estimation.

The derivative of the squared loss grows linearly with the residual. Intuitively, an observation whose residual is three times the typical size contributes roughly nine times as much to the squared-loss objective as a typical observation. In a long series this influence is diluted by many well-fitted observations, but a grey model estimated from only eight or nine balance equations offers no such dilution: a single outlying observation can substantially shift the estimated development coefficient and forcing term. Ridge regularization improves the conditioning of the inverse problem but does not bound this influence, because the loss gradient itself remains unbounded. The bounded exponential-squared loss has derivative $r\exp(-r^2/\sigma^2)$, which tends to zero for extreme residuals [36]. Incorporating this loss into grey RVFL estimation provides a mechanism for reducing the contribution of large residuals without changing the construction of the random features.

Randomized hidden nodes can also generate redundant or strongly correlated columns when the modeling window is short. Ridge regularization stabilizes the inverse problem but generally leaves all coefficients nonzero, whereas the ℓ_1 penalty produces sparse solutions whose magnitude depends on coefficient scaling. The ℓ_1/ℓ_2 norm ratio is scale invariant because $R(c\boldsymbol{\beta}) = \|c\boldsymbol{\beta}\|_1/\|c\boldsymbol{\beta}\|_2 = R(\boldsymbol{\beta})$ for every $c \neq 0$, and its proximal subproblem has been characterized analytically [37, 38]. Sparse grey RVFL models address hidden-feature redundancy but retain ordinary accumulation and

squared residual loss [35]. Conversely, the ℓ_1/ℓ_2-regularized extreme learning machine (ELM) with exponential-squared loss does not contain either a grey accumulated state or the direct branch that distinguishes RVFL from ELM [36]. A unified formulation is therefore needed to retain the direct RVFL link, penalize only the randomized hidden branch, and estimate the coefficients of both branches using a bounded-influence loss applied to an α-dependent grey residual.

These modeling choices are coupled through the data matrix. Changing α changes the accumulated response, background vector, standardized direct feature, and hidden-feature matrix; it consequently affects both the residual weights and the support selected for $\boldsymbol{\beta}$. Selecting the accumulation order first and then sparsifying a fixed design therefore need not yield the same solution as a nested procedure that optimizes the coupled hyperparameters around the coefficient-estimation problem. An additional internal validation split is also undesirable when the modeling sequence is extremely short: it reduces the number of available grey balance equations and reconstructs the CFA states, background values, and random-feature design on a different truncated sequence. The proposed hyperparameter optimization procedure therefore retains the complete modeling sequence and reserves the forecasting sequence exclusively for out-of-sample evaluation. The proposed formulation is termed the conformable fractional sparse robust grey RVFL model (CFSRGRVFL). In Table 1, 1-AGO denotes the first-order accumulated generating operation, CFUKRNGM denotes the conformable fractional-order unbiased kernel-regularized nonhomogeneous grey model, GRVFL denotes the grey RVFL model, and SKGRVFL denotes the sparse kernelized GRVFL model. The CFSRGRVFL solver uses an alternating direction method of multipliers (ADMM) framework with a difference-of-convex algorithm (DCA) update. Table 1 summarizes the methodological boundary. The contribution lies in the resulting partially penalized grey RVFL formulation and its blockwise derivation, rather than in merely placing existing components side by side.

Table 1. Comparison with the methods closest to the proposed formulation.

Method	Operator	Loss	Regularizer	Solver	Theoretical scope	Input–output structure
CFUKRNGM [18]	CFA	Squared	Kernel ℓ_2	Linear solve	Regularized kernel coefficient solution	Kernel forcing; no random-node bank
GRVFL [35]	1-AGO	Squared	Ridge ℓ_2	Closed form	Unique regularized least-squares solution	Direct input plus random hidden features
SKGRVFL [35]	1-AGO	Squared	Fractional ℓ_1 ratio	Coordinate descent	Monotone coordinate descent; no global forecasting guarantee	Kernel branch plus random hidden features
CFSRGRVFL	CFA and CFD	Bounded exponential-squared	ℓ_1/ℓ_2 on the hidden block	ADMM, norm-ratio proximal update, and DCA	Global ratio-proximal block solution and valid DC majorization; no full-sequence global claim	Unpenalized direct CFA-time link plus sparse random hidden features

These considerations motivate CFSRGRVFL. CFA and CFD control temporal weighting and restoration, while the RVFL forcing term combines a standardized CFA time input with randomized hidden features. A bounded exponential-squared loss is used for coefficient estimation, and the ℓ_1/ℓ_2 norm-ratio regularizer is imposed only on the hidden-output coefficients so that hidden features can be selected without removing the direct RVFL branch. The resulting objective is non-smooth and

non-convex. Its splitting and norm-ratio update draw on the bounded-loss ELM formulation and the ℓ_1/ℓ_2 proximal characterization [36,38]. Outside this inner estimation problem, Bayesian optimization minimizes an objective averaged over a fixed set of random seeds, thereby optimizing the coupled hyperparameters under a finite evaluation budget while retaining the complete modeling sequence. The bounded-loss scale is parameterized relative to the transformed-response dispersion [35, 39]. The primary contributions of this paper are summarized below:

- CFSRGRVFL integrates CFA and CFD with a complete grey RVFL structure to model order-dependent temporal weighting and nonlinear features within a unified framework.
- A bounded-loss, partially sparse coefficient estimator retains the direct RVFL link while selecting informative random hidden features.
- An ADMM scheme with a DCA update is developed with explicit subproblem updates and a practical support-size search for the ℓ_1/ℓ_2 proximal step. A Bayesian hyperparameter optimization procedure uses a dimensionless multiplier for the bounded-loss scale and a fixed set of random seeds to jointly search the fractional order, hidden-layer size, regularization parameter, ADMM penalty parameter, and hidden-layer scale parameters without further partitioning the modeling sequence.
- Experiments on five energy time series compare CFSRGRVFL with a residual learning-based grey system model (GM-ResNet), a grey ELM-type model (GM-ELM), and a neural grey system model with Bayesian regularization (NGSM) across nine metrics and demonstrate its overall forecasting advantage within the tested settings.

The rest of this paper is organized as follows: Section 2 reviews random-feature neural grey and GRVFL models. Section 3 presents the proposed CFSRGRVFL model. Section 4 develops the ADMM-based algorithm with a DCA update. Section 5 presents the Bayesian hyperparameter optimization procedure for short modeling sequences. Section 6 reports the numerical experiments, and Section 7 concludes this work.

2. Random-feature neural grey and grey RVFL models

Ma et al. [29] introduced an evolutionary feature-based neural grey model and referred to it as GreyRVFL. Its forcing term, however, contains only randomized hidden responses and omits the direct input–output link that defines a complete RVFL network. Gong and Zong [35] therefore distinguish this formulation as GM-ELM model and subsequently construct the GRVFL model by concatenating direct input features and randomized hidden features. The distinction is structural: A GM-ELM-type forcing term contains only randomized hidden features, whereas a complete RVFL structure additionally contains a direct (skip-layer) input–output link alongside the randomized hidden layer. Following this distinction, the model of Ma et al. is referred to as GM-ELM hereafter. This section reviews the common grey framework and then separates these two network structures.

Let $N \geq 2$ be the number of observations and let $d_{\text{in}} \geq 1$ be the input dimension. For $k = 1, \ldots, N$, let $T^{(0)}(k) \in \mathbb{R}$ be the target value and let

$$\boldsymbol{I}^{(0)}(k) = \left(I_1^{(0)}(k), I_2^{(0)}(k), \ldots, I_{d_{\text{in}}}^{(0)}(k)\right)^{\top} \in \mathbb{R}^{d_{\text{in}}}$$

be the input vector. Their first-order accumulated sequences are

$$T^{(1)}(k) = \sum_{\iota=1}^{k} T^{(0)}(\iota), \qquad I_i^{(1)}(k) = \sum_{\iota=1}^{k} I_i^{(0)}(\iota), \quad i = 1, \ldots, d_{\text{in}},$$

with $\boldsymbol{I}^{(1)}(k) = (I_1^{(1)}(k), \ldots, I_{d_{\text{in}}}^{(1)}(k))^\top$. Let $\widetilde{\boldsymbol{I}}^{(1)} : [1, N] \to \mathbb{R}^{d_{\text{in}}}$ denote the piecewise-linear interpolant of the accumulated input sequence. The common whitening equation is

$$\frac{dT^{(1)}(t)}{dt} + aT^{(1)}(t) = f\left(\widetilde{\boldsymbol{I}}^{(1)}(t)\right), \tag{2.1}$$

where $a \in \mathbb{R}$ is the development coefficient. For $k = 2, \ldots, N$, define

$$z^{(1)}(k) = \frac{T^{(1)}(k-1) + T^{(1)}(k)}{2}, \qquad \boldsymbol{v}(k) = \frac{\boldsymbol{I}^{(1)}(k-1) + \boldsymbol{I}^{(1)}(k)}{2}.$$

Integration over $[k-1, k]$, followed by the standard trapezoidal and midpoint approximations, gives the grey balance equation

$$T^{(0)}(k) + az^{(1)}(k) = f\left(\boldsymbol{v}(k)\right). \tag{2.2}$$

Both reviewed models use the sigmoid activation

$$S(\boldsymbol{x}; \boldsymbol{w}, b) = \frac{1}{1 + \exp[-(\boldsymbol{x}^\top \boldsymbol{w} + b)]}.$$

2.1. The GreyRVFL/GM-ELM formulation

In the model of Ma et al. [29], the forcing term is represented only by L randomized hidden responses:

$$f_{\text{M}}\left(\widetilde{\boldsymbol{I}}^{(1)}(t)\right) = \sum_{j=1}^{L} \beta_j S\left(\widetilde{\boldsymbol{I}}^{(1)}(t); \boldsymbol{w}_j, b_j\right), \tag{2.3}$$

where $\boldsymbol{w}_j \in \mathbb{R}^{d_{\text{in}}}$ and $b_j \in \mathbb{R}$ are hidden parameters and $\beta_j \in \mathbb{R}$ is an output coefficient. Substitution into (2.2) gives

$$T^{(0)}(k) + az^{(1)}(k) = \sum_{j=1}^{L} \beta_j S\left(\boldsymbol{v}(k); \boldsymbol{w}_j, b_j\right). \tag{2.4}$$

Thus, despite the terminology used in the original article, (2.3) has an ELM-type hidden-feature structure rather than a complete RVFL structure.

To summarize its coefficient estimation, define

$$\boldsymbol{y} = \left(T^{(0)}(2), \ldots, T^{(0)}(N)\right)^\top, \qquad \boldsymbol{z}^{(1)} = \left(z^{(1)}(2), \ldots, z^{(1)}(N)\right)^\top,$$

and let $\boldsymbol{H}_{\text{M}} \in \mathbb{R}^{(N-1)\times L}$ have entries

$$(\boldsymbol{H}_{\text{M}})_{k-1,j} = S\left(\boldsymbol{v}(k); \boldsymbol{w}_j, b_j\right).$$

With

$$\boldsymbol{U}_{\text{M}} = \left[-\boldsymbol{z}^{(1)} \quad \boldsymbol{H}_{\text{M}}\right], \qquad \boldsymbol{\vartheta}_{\text{M}} = (a, \boldsymbol{\beta}^\top)^\top,$$

Ma et al. [29] estimates the development coefficient and output weights by

$$\boldsymbol{\vartheta}_{\mathrm{M}}^{*} \in \arg\min_{\boldsymbol{\vartheta}_{\mathrm{M}}} \|\boldsymbol{y} - \boldsymbol{U}_{\mathrm{M}}\boldsymbol{\vartheta}_{\mathrm{M}}\|_2^2 . \tag{2.5}$$

When $\boldsymbol{U}_{\mathrm{M}}^{\top}\boldsymbol{U}_{\mathrm{M}}$ is nonsingular, the stated least-squares solution is

$$\boldsymbol{\vartheta}_{\mathrm{M}}^{*} = \left(\boldsymbol{U}_{\mathrm{M}}^{\top}\boldsymbol{U}_{\mathrm{M}}\right)^{-1}\boldsymbol{U}_{\mathrm{M}}^{\top}\boldsymbol{y}.$$

The hidden weights and biases are further optimized by evolutionary algorithms using a mean absolute percentage error (MAPE) objective, whereas the number of hidden nodes is selected by grid search on validation data [29].

2.2. The GRVFL formulation

Gong and Zong [35] restore the defining RVFL direct link. For $k = 2, \ldots, N$, they construct the adjacent-state input

$$\boldsymbol{s}_k^{(1)} = \begin{bmatrix} \boldsymbol{I}^{(1)}(k-1) \\ \boldsymbol{I}^{(1)}(k) \end{bmatrix} \in \mathbb{R}^{2d_{\mathrm{in}}}.$$

Let $\widetilde{\boldsymbol{w}}_j \in \mathbb{R}^{2d_{\mathrm{in}}}$ and $\widetilde{b}_j \in \mathbb{R}$ be random hidden parameters. The complete GRVFL feature row is

$$\boldsymbol{h}_k = \left[(\boldsymbol{s}_k^{(1)})^{\top},\, S(\boldsymbol{s}_k^{(1)}; \widetilde{\boldsymbol{w}}_1, \widetilde{b}_1), \ldots, S(\boldsymbol{s}_k^{(1)}; \widetilde{\boldsymbol{w}}_L, \widetilde{b}_L)\right] \in \mathbb{R}^{2d_{\mathrm{in}}+L}. \tag{2.6}$$

For an output-weight vector $\boldsymbol{\theta}_{\mathrm{G}} \in \mathbb{R}^{2d_{\mathrm{in}}+L}$, the discrete GRVFL equation is

$$T^{(0)}(k) + az^{(1)}(k) = \boldsymbol{h}_k\boldsymbol{\theta}_{\mathrm{G}}. \tag{2.7}$$

The first $2d_{\mathrm{in}}$ entries of $\boldsymbol{h}_k$ form the direct input block, whereas the final L entries form the randomized hidden block.

Let $\boldsymbol{H}_{\mathrm{G}} \in \mathbb{R}^{(N-1)\times(2d_{\mathrm{in}}+L)}$ have row $k-1$ equal to $\boldsymbol{h}_k$, and define

$$\overline{\boldsymbol{H}}_{\mathrm{G}} = \begin{bmatrix} \boldsymbol{H}_{\mathrm{G}} & -\boldsymbol{z}^{(1)} \end{bmatrix}, \qquad \overline{\boldsymbol{\theta}}_{\mathrm{G}} = \begin{bmatrix} \boldsymbol{\theta}_{\mathrm{G}} \\ a \end{bmatrix}.$$

Gong and Zong [35] estimate the augmented coefficient vector using Tikhonov-regularized least squares:

$$\min_{\overline{\boldsymbol{\theta}}_{\mathrm{G}}} \frac{1}{2}\left\|\overline{\boldsymbol{H}}_{\mathrm{G}}\overline{\boldsymbol{\theta}}_{\mathrm{G}} - \boldsymbol{y}\right\|_2^2 + \frac{\rho_{\mathrm{G}}}{2}\left\|\overline{\boldsymbol{\theta}}_{\mathrm{G}}\right\|_2^2, \qquad \rho_{\mathrm{G}} > 0. \tag{2.8}$$

The unique solution is

$$\overline{\boldsymbol{\theta}}_{\mathrm{G}}^{*} = \left(\overline{\boldsymbol{H}}_{\mathrm{G}}^{\top}\overline{\boldsymbol{H}}_{\mathrm{G}} + \rho_{\mathrm{G}}\,\mathrm{Id}_{2d_{\mathrm{in}}+L+1}\right)^{-1}\overline{\boldsymbol{H}}_{\mathrm{G}}^{\top}\boldsymbol{y}.$$

The positive ridge parameter removes the full-column-rank requirement imposed by the unregularized inverse in (2.5).

2.3. Time response and relation to the proposed model

Both articles obtain the accumulated response from the variation-of-constants formula associated with (2.1). Writing $q(\iota)$ for the forcing assigned to the interval $[\iota-1,\iota]$ and using the interval index consistently, the midpoint response is

$$\widehat{T}^{(1)}(k)=T^{(0)}(1)\exp[-a(k-1)]+\sum_{\iota=2}^{k}\exp\left[-a\left(k-\iota+\frac{1}{2}\right)\right]q(\iota),\qquad k=2,\dots,N, \tag{2.9}$$

where

$$q(\iota)=\begin{cases}\displaystyle\sum_{j=1}^{L}\beta_j S\left(\boldsymbol{v}(\iota);\boldsymbol{w}_j,b_j\right), & \text{for the GM-ELM formulation,}\\ \boldsymbol{h}_\iota\boldsymbol{\theta}_{\mathrm{G}}, & \text{for the GRVFL formulation.}\end{cases}$$

The original-scale response is recovered by

$$\widehat{T}^{(0)}(k)=\widehat{T}^{(1)}(k)-\widehat{T}^{(1)}(k-1).$$

The proposed model retains the complete RVFL principle of combining a direct link with random hidden features, but it does not reproduce the adjacent-state input of GRVFL. Instead, it uses a standardized CFA time-background feature as both the direct input and the argument of the hidden mapping. It also replaces ordinary accumulation and squared-loss estimation with CFA/CFD and a bounded robust sparse criterion whose ℓ_1/ℓ_2 penalty acts only on the hidden-output coefficients.

3. The proposed CFSRGRVFL model

The proposed model is presented in the standard order used in grey forecasting theory. The accumulation and restoration operators are introduced first, followed by the whitening equation, the time response formula, and the robust sparse coefficient-estimation model.

3.1. Definitions of CFA and CFD

The classical first-order accumulated generating operation uses fixed weights. In the proposed model, this operator is replaced by conformable fractional accumulation with an adjustable accumulation order. The corresponding inverse operation is the conformable fractional difference. Before the operators are defined, Table 2 summarizes the main notation used throughout the paper; in particular, it fixes the correspondence between an original-scale sequence $\boldsymbol{x}^{(0)}$ and its CFA image $\boldsymbol{x}^{(\alpha)}=\nabla^{\alpha}\boldsymbol{x}^{(0)}$.

Table 2. Main notation used throughout the paper.

Symbol	Meaning
$\boldsymbol{x}^{(0)}$, $x^{(0)}(k)$	generic original-scale sequence and its value at index k
∇^{α}	CFA operator of order $\alpha \in (0,1]$, with $\boldsymbol{x}^{(\alpha)} = \nabla^{\alpha}\boldsymbol{x}^{(0)}$
$\boldsymbol{x}^{(\alpha)}$, $x^{(\alpha)}(k)$	CFA-accumulated sequence and its value at index k
Δ^{α}	CFD operator, satisfying $\Delta^{\alpha}\nabla^{\alpha}\boldsymbol{x}^{(0)} = \boldsymbol{x}^{(0)}$
$\boldsymbol{T}^{(0)}$, $\boldsymbol{T}^{(\alpha)}$	target sequence and its CFA sequence
$\tau^{(0)}(k)$, $\tau^{(\alpha)}(k)$	normalized time input and its CFA value
$z^{(\alpha)}(k)$	background value of $\boldsymbol{T}^{(\alpha)}$
$\bar{v}_{\tau}^{(\alpha)}(k)$	standardized background time input
a, γ, $\boldsymbol{\beta}$	development coefficient, direct-link coefficient, and hidden-output vector
$\boldsymbol{H}$, $\boldsymbol{P}$	hidden-feature matrix and complete RVFL feature matrix
ϕ_{σ}, σ, κ_{σ}	bounded exponential-squared loss, its scale, and the dimensionless multiplier
$R(\boldsymbol{\beta})$, λ	ℓ_1/ℓ_2 norm-ratio regularizer and its weight
η, $\boldsymbol{\mu}$	ADMM penalty parameter and multiplier vector
N, N_{tr}, n_{ev}	sequence length, modeling-sample size, and evaluation-window length
$\bar{x}$	mean of the observed values over the evaluated window

The following notation is used in both definitions. Let $N \geq 2$ be the sequence length and let $\mathcal{K}_N = \{1, 2, \ldots, N\}$ be the index set. For a finite sequence $\boldsymbol{x}^{(0)} = (x^{(0)}(1), \ldots, x^{(0)}(N))^{\top}$, the superscript (0) denotes the original scale, and $x^{(0)}(\iota)$ is the value at index $\iota \in \mathcal{K}_N$. Although the target sequence in grey forecasting is usually nonnegative, the operators below are stated for real-valued sequences because the same CFA notation is also used for the normalized time input.

Definition 3.1 (CFA)**.** For $\alpha \in (0,1]$ and $\boldsymbol{x}^{(0)} \in \mathbb{R}^N$, the conformable fractional accumulation operator $\nabla^{\alpha} : \mathbb{R}^N \to \mathbb{R}^N$ generates $\boldsymbol{x}^{(\alpha)} = \nabla^{\alpha}\boldsymbol{x}^{(0)}$ by

$$x^{(\alpha)}(k) = \left(\nabla^{\alpha}\boldsymbol{x}^{(0)}\right)(k) = \sum_{\iota=1}^{k} \iota^{\alpha-1} x^{(0)}(\iota), \qquad k = 1, \ldots, N. \tag{3.1}$$

In (3.1), ι is only the summation index, and $\iota^{\alpha-1}$ is the conformable fractional weight assigned to $x^{(0)}(\iota)$ before cumulative summation. When $\alpha = 1$, (3.1) becomes the ordinary first-order accumulated generating operation. When $0 < \alpha < 1$, this factor changes the weight of each observation before accumulation. Throughout the paper, the superscript (0) always denotes a sequence on the original scale, and the superscript (α) denotes its image under the CFA operator; the operator notation $\nabla^{\alpha}\boldsymbol{x}^{(0)}$ and the superscript notation $\boldsymbol{x}^{(\alpha)}$ in (3.1) therefore refer to the same transformation.

After an accumulated sequence $\boldsymbol{x}^{(\alpha)}$ has been obtained, restoration to the original scale is described by the inverse sequence operator. The first component is retained, and the remaining components are computed from adjacent accumulated values.

Definition 3.2 (CFD)**.** For $\alpha \in (0,1]$ and $\boldsymbol{x}^{(\alpha)} \in \mathbb{R}^N$, the conformable fractional difference operator

$\Delta^\alpha : \mathbb{R}^N \to \mathbb{R}^N$ generates $\boldsymbol{x}^{(0)} = \Delta^\alpha \boldsymbol{x}^{(\alpha)}$ by

$$x^{(0)}(k) = \left(\Delta^\alpha \boldsymbol{x}^{(\alpha)}\right)(k) = \begin{cases} x^{(\alpha)}(1), & k = 1, \\ k^{1-\alpha}\left(x^{(\alpha)}(k) - x^{(\alpha)}(k-1)\right), & k = 2, \ldots, N. \end{cases} \tag{3.2}$$

For $k \geq 2$, the factor $k^{1-\alpha}$ reverses the accumulation weight $k^{\alpha-1}$ in (3.1), so $(\Delta^\alpha \nabla^\alpha \boldsymbol{x}^{(0)})(k) = x^{(0)}(k)$ for $k = 1, \ldots, N$. After an accumulated sequence is fitted, the original-scale sequence is recovered by

$$\widehat{x}^{(0)}(k) = \begin{cases} \widehat{x}^{(\alpha)}(1), & k = 1, \\ k^{1-\alpha}\left(\widehat{x}^{(\alpha)}(k) - \widehat{x}^{(\alpha)}(k-1)\right), & k = 2, \ldots, N. \end{cases} \tag{3.3}$$

The weighting and cumulative-summation mechanism of the CFA operator in (3.1) is illustrated in Figure 1. The compact block represents the mapping from the weighted term $k^{\alpha-1}T^{(0)}(k)$ to the accumulated value $T^{(\alpha)}(k)$, whereas the unfolded form shows that $T^{(\alpha)}(k)$ is obtained by adding all weighted observations up to index k.

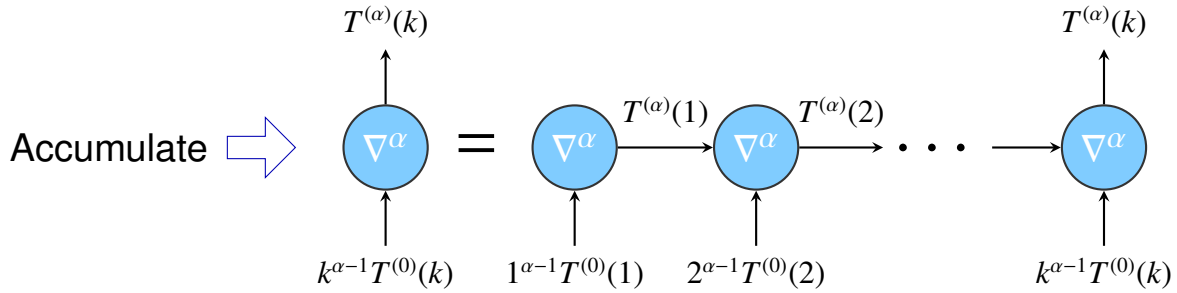


Figure 1. The structure of conformable fractional accumulation.

3.2. CFSRGRVFL whitening equation and time response function

Let $\boldsymbol{T}^{(0)} = (T^{(0)}(1), \ldots, T^{(0)}(N))^\top \in \mathbb{R}^N_{\geq 0}$ be the target sequence, where $N \geq 2$, and let $\boldsymbol{T}^{(\alpha)} = \nabla^\alpha \boldsymbol{T}^{(0)}$ be its CFA sequence. The model is specified without exogenous inputs. Fix a normalization length $N_s \geq 2$; in model fitting set $N_s = N$ and retain this value for forecasting. For every integer $k \geq 1$, define the normalized time input and its CFA value by

$$\tau^{(0)}(k) = \frac{k-1}{N_s - 1}, \qquad \tau^{(\alpha)}(k) = \sum_{\iota=1}^{k} \iota^{\alpha-1}\tau^{(0)}(\iota).$$

Only the values $k = 1, \ldots, N$ are used for fitting; the same formula defines future time inputs without re-estimating N_s. For $k = 2, \ldots, N$, define the background values and the standardized background time input by

$$z^{(\alpha)}(k) = \frac{T^{(\alpha)}(k-1) + T^{(\alpha)}(k)}{2}, \qquad v_\tau^{(\alpha)}(k) = \frac{\tau^{(\alpha)}(k-1) + \tau^{(\alpha)}(k)}{2},$$

and

$$\bar{v}_\tau^{(\alpha)}(k) = \frac{v_\tau^{(\alpha)}(k) - \mu_v}{s_v}.$$

The training mean and scale are defined explicitly by

$$\mu_v = \frac{1}{N-1}\sum_{k=2}^{N} v_\tau^{(\alpha)}(k), \qquad s_{v,0} = \left[\frac{1}{N-1}\sum_{k=2}^{N}\left(v_\tau^{(\alpha)}(k)-\mu_v\right)^2\right]^{1/2},$$

with $s_v = s_{v,0}$ if $s_{v,0} > 0$ and $s_v = 1$ otherwise. Thus, the standardization is well defined even when $N = 2$. The same training values (μ_v, s_v) are retained during forecasting. To connect the discrete CFA input to the continuous whitening equation, let $\widetilde{\tau}^{(\alpha)} : [1,\infty) \to \mathbb{R}$ be the piecewise-linear interpolant satisfying $\widetilde{\tau}^{(\alpha)}(k) = \tau^{(\alpha)}(k)$ at every integer $k \geq 1$, and define

$$\bar{\tau}^{(\alpha)}(t) = \frac{\widetilde{\tau}^{(\alpha)}(t)-\mu_v}{s_v}.$$

In particular, $\bar{\tau}^{(\alpha)}(k-1/2) = \bar{v}_\tau^{(\alpha)}(k)$. The accumulated target is assumed to satisfy the whitening equation

$$\frac{dT^{(\alpha)}(t)}{dt} + aT^{(\alpha)}(t) = f_{\gamma,\boldsymbol{\beta}}(\bar{\tau}^{(\alpha)}(t)), \tag{3.4}$$

where $a \in \mathbb{R}$ is the grey development coefficient and $f_{\gamma,\boldsymbol{\beta}} : \mathbb{R} \to \mathbb{R}$ is the RVFL driving term.

Let $L \geq 1$ be the number of hidden nodes. For $j = 1,\ldots,L$, scalar random hidden parameters $w_j, b_j \in \mathbb{R}$ are generated by

$$w_j \sim \mathcal{N}(0, c_w^2), \qquad b_j \sim \mathrm{Unif}[-c_b, c_b],$$

where $c_w, c_b > 0$ are hidden-layer scale parameters. Let $\gamma \in \mathbb{R}$ be the coefficient of the direct standardized-time input and let $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_L)^\top$ contain the hidden-output coefficients. The continuous RVFL driving function is

$$f_{\gamma,\boldsymbol{\beta}}(x) = \gamma x + \sum_{j=1}^{L}\beta_j S(w_j x + b_j), \qquad x \in \mathbb{R}. \tag{3.5}$$

The function in (3.5) closes the whitening equation in (3.4). At the midpoint of the k-th unit interval, define the hidden responses and interval driving value for every integer $k \geq 2$ by

$$h_j(k) = S(\bar{v}_\tau^{(\alpha)}(k)w_j + b_j), \qquad q(k) = \gamma\bar{v}_\tau^{(\alpha)}(k) + \sum_{j=1}^{L}\beta_j h_j(k).$$

Equivalently, $q(k) = f_{\gamma,\boldsymbol{\beta}}(\bar{v}_\tau^{(\alpha)}(k))$. The activation function S is the sigmoid function introduced in Section 2. The first term in $q(k)$ is the direct input–output connection, whereas the summation is the randomized nonlinear branch. Their concatenation is the standard RVFL representation.

Integrating (3.4) over the unit interval $[k-1,k]$ first gives the exact relation

$$T^{(\alpha)}(k) - T^{(\alpha)}(k-1) + a\int_{k-1}^{k} T^{(\alpha)}(t)\,dt = \int_{k-1}^{k} f_{\gamma,\boldsymbol{\beta}}\left(\bar{\tau}^{(\alpha)}(t)\right)\,dt. \tag{3.6}$$

The state integral is approximated by the trapezoidal background value, and the forcing integral is evaluated by the midpoint rule:

$$\int_{k-1}^{k} T^{(\alpha)}(t)\,dt \approx z^{(\alpha)}(k), \qquad \int_{k-1}^{k} f_{\gamma,\boldsymbol{\beta}}(\bar{\tau}^{(\alpha)}(t))\,dt \approx q(k).$$

Moreover, the CFA definition and its CFD inverse imply

$$T^{(\alpha)}(k) - T^{(\alpha)}(k-1) = k^{\alpha-1}T^{(0)}(k) = \frac{T^{(0)}(k)}{k^{1-\alpha}}. \tag{3.7}$$

Substitution of these two quadrature approximations and (3.7) into (3.6) leads to the estimating equation, for $k = 2, \ldots, N$,

$$\frac{T^{(0)}(k)}{k^{1-\alpha}} + az^{(\alpha)}(k) = q(k) = \gamma\bar{v}_{\tau}^{(\alpha)}(k) + \sum_{j=1}^{L}\beta_j h_j(k). \tag{3.8}$$

Equation (3.8) is the discrete grey balance equation used to estimate the unknown coefficients. Following the variation-of-constants construction used for the grey RVFL response by Gong and Zong [35], the forecasting response is derived directly from the continuous whitening Eq (3.4). Multiplication by the integrating factor $\exp(at)$ gives

$$\frac{d}{dt}\left[\exp(at)T^{(\alpha)}(t)\right] = \exp(at)f_{\gamma,\boldsymbol{\beta}}\left(\bar{\tau}^{(\alpha)}(t)\right).$$

Because $T^{(\alpha)}(1) = T^{(0)}(1)$, integration from 1 to t yields

$$T^{(\alpha)}(t) = T^{(0)}(1)\exp[-a(t-1)] + \int_1^t \exp[-a(t-s)]f_{\gamma,\boldsymbol{\beta}}\left(\bar{\tau}^{(\alpha)}(s)\right)ds. \tag{3.9}$$

For an integer response point $k \geq 2$, the integral in (3.9) is split over the unit intervals $[\iota-1, \iota]$, $\iota = 2, \ldots, k$. On each interval, the midpoint rule, together with $\bar{\tau}^{(\alpha)}(\iota - 1/2) = \bar{v}_{\tau}^{(\alpha)}(\iota)$, gives

$$\int_{\iota-1}^{\iota} \exp[-a(k-s)]f_{\gamma,\boldsymbol{\beta}}\left(\bar{\tau}^{(\alpha)}(s)\right)ds \approx \exp\left[-a\left(k-\iota+\frac{1}{2}\right)\right]q(\iota).$$

Since each interval has unit length, no additional multiplicative factor is introduced by the midpoint rule. Substitution into (3.9) gives the CFSRGRVFL time response sequence

$$\widehat{T}^{(\alpha)}(k) = T^{(0)}(1)\exp[-a(k-1)] + \sum_{\iota=2}^{k}\exp\left[-a\left(k-\iota+\frac{1}{2}\right)\right]q(\iota), \qquad k \geq 2. \tag{3.10}$$

With $\widehat{T}^{(\alpha)}(1) = T^{(0)}(1)$, multiplying the response at $k-1$ by $\exp(-a)$ and subtracting it from the response at k cancels every interval term except the one indexed by k. Hence,

$$\widehat{T}^{(\alpha)}(k) - \exp(-a)\widehat{T}^{(\alpha)}(k-1) = \exp\left(-\frac{a}{2}\right)q(k),$$

and the finite-sum response in (3.10) has the equivalent one-step form

$$\widehat{T}^{(\alpha)}(k) = \exp(-a)\widehat{T}^{(\alpha)}(k-1) + \exp\left(-\frac{a}{2}\right)q(k), \qquad k \geq 2. \tag{3.11}$$

The fitted or forecast values on the original scale are recovered from (3.10), or equivalently (3.11), by the CFD restoration formula (3.3). Both the estimating equation and the response sequence originate from (3.4), but they use different quadrature constructions: (3.8) approximates the state and forcing integrals in the integrated balance equation, whereas (3.10) applies the midpoint rule after the exponential kernel has been introduced by variation of constants. Therefore, they are linked through the same continuous model but are not algebraically identical discrete equations. This distinction follows the grey time-response construction reviewed in Section 2.

3.3. Parameter estimation

The parameters a, γ, and $\boldsymbol{\beta}$ in (3.8) are estimated after the CFA order, hidden-layer size, random-feature scales, and random seed have been fixed. In the original grey RVFL model, this step is a least-squares estimation problem. The present model replaces it with a robust sparse optimization formulation. The direct-link coefficient γ is excluded from the sparsity penalty, so the direct branch remains available and cannot be deleted by hidden-node selection; its fitted value may nevertheless be zero. The ℓ_1/ℓ_2 penalty is applied only to $\boldsymbol{\beta}$ because its purpose is to remove redundant random hidden nodes.

For compact matrix notation, define the response vector $\boldsymbol{y}^{(\alpha)} \in \mathbb{R}^{N-1}$ as

$$\boldsymbol{y}^{(\alpha)} = \left(\frac{T^{(0)}(2)}{2^{1-\alpha}}, \frac{T^{(0)}(3)}{3^{1-\alpha}}, \ldots, \frac{T^{(0)}(N)}{N^{1-\alpha}} \right)^{\top}, \tag{3.12}$$

and define the background vector $\boldsymbol{z}^{(\alpha)} \in \mathbb{R}^{N-1}$ as

$$\boldsymbol{z}^{(\alpha)} = \left(z^{(\alpha)}(2), z^{(\alpha)}(3), \ldots, z^{(\alpha)}(N) \right)^{\top}. \tag{3.13}$$

The direct-input feature vector is

$$\boldsymbol{v}^{(\alpha)} = \left(\bar{v}_{\tau}^{(\alpha)}(2), \bar{v}_{\tau}^{(\alpha)}(3), \ldots, \bar{v}_{\tau}^{(\alpha)}(N) \right)^{\top} \in \mathbb{R}^{N-1}. \tag{3.14}$$

The hidden-feature matrix $\boldsymbol{H} \in \mathbb{R}^{(N-1)\times L}$ is

$$\boldsymbol{H} = \begin{bmatrix} h_1(2) & h_2(2) & \cdots & h_L(2) \\ h_1(3) & h_2(3) & \cdots & h_L(3) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(N) & h_2(N) & \cdots & h_L(N) \end{bmatrix}. \tag{3.15}$$

Consequently, the complete RVFL feature matrix and its output-coefficient vector are

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{v}^{(\alpha)} & \boldsymbol{H} \end{bmatrix} \in \mathbb{R}^{(N-1)\times(L+1)}, \qquad \boldsymbol{\omega} = \begin{bmatrix} \gamma \\ \boldsymbol{\beta} \end{bmatrix} \in \mathbb{R}^{L+1}. \tag{3.16}$$

The first column of $\boldsymbol{P}$ is the direct link and the remaining L columns are hidden responses. To include the grey development term in the same linear representation, define

$$\boldsymbol{X}_{\alpha} = \begin{bmatrix} -\boldsymbol{z}^{(\alpha)} & \boldsymbol{v}^{(\alpha)} & \boldsymbol{H} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{z}^{(\alpha)} & \boldsymbol{P} \end{bmatrix} \in \mathbb{R}^{(N-1)\times(L+2)}, \qquad \boldsymbol{\theta} = \begin{bmatrix} a \\ \boldsymbol{\omega} \end{bmatrix} = \begin{bmatrix} a \\ \gamma \\ \boldsymbol{\beta} \end{bmatrix} \in \mathbb{R}^{L+2}. \tag{3.17}$$

With these definitions, the k-th row of $\boldsymbol{y}^{(\alpha)} - \boldsymbol{X}_{\alpha}\boldsymbol{\theta}$ is the left-hand side of (3.8) minus its righthand side. The resulting residual vector is

$$\boldsymbol{r}(a, \gamma, \boldsymbol{\beta}) = \boldsymbol{y}^{(\alpha)} + a\boldsymbol{z}^{(\alpha)} - \gamma\boldsymbol{v}^{(\alpha)} - \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{y}^{(\alpha)} + a\boldsymbol{z}^{(\alpha)} - \boldsymbol{P}\boldsymbol{\omega} = \boldsymbol{y}^{(\alpha)} - \boldsymbol{X}_{\alpha}\boldsymbol{\theta}. \tag{3.18}$$

This residual vector lies in $\mathbb{R}^{N-1}$ and links the grey dynamic equation to the coefficient-estimation objective.

For $\sigma > 0$, define the bounded exponential-squared loss by

$$\phi_\sigma(r) = \frac{\sigma^2}{2}\left(1 - \exp\left(-\frac{r^2}{\sigma^2}\right)\right). \tag{3.19}$$

Its local expansion and first two derivatives are

$$\phi_\sigma(r) = \frac{r^2}{2} - \frac{r^4}{4\sigma^2} + O(r^6), \qquad r \to 0, \tag{3.20a}$$

$$\phi'_\sigma(r) = r\exp\left(-\frac{r^2}{\sigma^2}\right), \tag{3.20b}$$

$$\phi''_\sigma(r) = \exp\left(-\frac{r^2}{\sigma^2}\right)\left(1 - \frac{2r^2}{\sigma^2}\right). \tag{3.20c}$$

Thus, the loss behaves quadratically near the origin, satisfies

$$0 \le \phi_\sigma(r) < \frac{\sigma^2}{2}, \qquad \max_{r\in\mathbb{R}} |\phi'_\sigma(r)| = \frac{\sigma}{\sqrt{2e}},$$

and has a derivative that tends to zero as $|r| \to \infty$. The second derivative in (3.20c) changes sign at $|r| = \sigma/\sqrt{2}$, so the loss is locally convex and becomes concave in its tails; this also shows directly why the loss is non-convex.

Redundant hidden nodes are penalized through the ℓ_1/ℓ_2 ratio. Following the lower-semicontinuous extension used in the proximal analysis of Tao [38], define

$$R(\boldsymbol{\beta}) = \begin{cases} \dfrac{\|\boldsymbol{\beta}\|_1}{\|\boldsymbol{\beta}\|_2}, & \boldsymbol{\beta} \neq \mathbf{0}, \\ 1, & \boldsymbol{\beta} = \mathbf{0}. \end{cases} \tag{3.21}$$

For every nonzero $\boldsymbol{\beta} \in \mathbb{R}^L$ with $s_\beta = \|\boldsymbol{\beta}\|_0$, the norm inequalities give

$$1 \le R(\boldsymbol{\beta}) = \frac{\|\boldsymbol{\beta}\|_1}{\|\boldsymbol{\beta}\|_2} \le \sqrt{s_\beta} \le \sqrt{L}. \tag{3.22}$$

The lower bound is attained by a one-sparse vector, whereas the upper bound $\sqrt{s_\beta}$ is attained when the s_β nonzero entries have equal magnitude. Minimizing the ratio therefore favors concentration on a smaller hidden-node support. Moreover,

$$R(c\boldsymbol{\beta}) = R(\boldsymbol{\beta}), \qquad c \neq 0,$$

so the regularizer controls the coefficient direction and support rather than its overall magnitude. Assigning $R(\mathbf{0}) = 1$ makes R lower semicontinuous on $\mathbb{R}^L$, although it remains discontinuous at the origin along most directions. This extension permits the proximal subproblem to be stated on the full Euclidean space and removes the need for an artificial nonzero safeguard.

Define the sample loss by

$$\Phi_\sigma(\boldsymbol{r}) = \sum_{i=1}^{N-1} \phi_\sigma(r_i), \qquad \boldsymbol{r} \in \mathbb{R}^{N-1}.$$

Combining (3.18), (3.19), and (3.21) yields the coefficient-estimation model

$$\min_{a,\gamma\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^L} \Phi_\sigma\left(\boldsymbol{y}^{(\alpha)} + a\boldsymbol{z}^{(\alpha)} - \gamma\boldsymbol{v}^{(\alpha)} - \boldsymbol{H}\boldsymbol{\beta}\right) + \lambda R(\boldsymbol{\beta}). \tag{3.23}$$

Here $\lambda > 0$ controls the norm-ratio term. For fixed values of $(\alpha, L, c_w, c_b, \lambda, \sigma, \eta)$ and a fixed random seed, (3.23), together with its ADMM–DCA solver, is the inner problem that estimates only $(a, \gamma, \boldsymbol{\beta})$. Section 5 defines a dimensionless loss-scale factor, converts it to σ on the CFA-response scale, and specifies the outer optimization of the coupled model and algorithmic hyperparameters.

The optimization problem (3.23) is non-smooth and non-convex because it contains the norm-ratio regularizer $R(\boldsymbol{\beta})$ and the bounded exponential-squared loss. The direct coefficient γ is estimated jointly with a and $\boldsymbol{\beta}$ but is excluded from R, so zeroing entries of $\boldsymbol{\beta}$ removes random hidden nodes without deleting the defining RVFL skip connection. The explicit residual representation in (3.17) and (3.18) also identifies the two blocks used by the solver: (a, γ) remain unpenalized, whereas only the hidden-output block $\boldsymbol{\beta}$ is passed to the norm-ratio proximal step. Section 4 presents this solution method.

Figure 2 summarizes the architecture of the proposed CFSRGRVFL model.

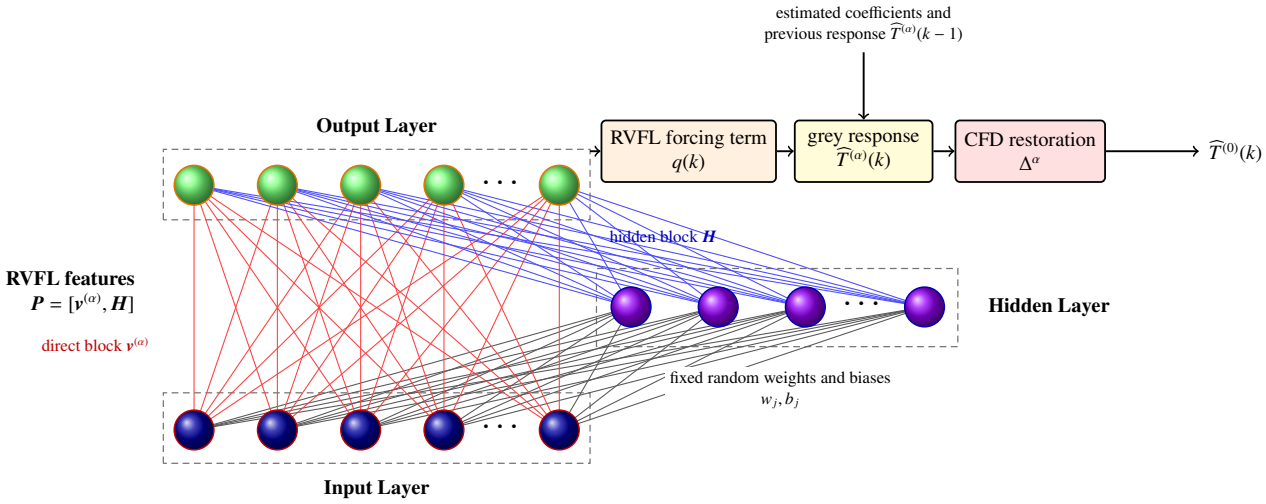


Figure 2. Architecture of CFSRGRVFL. The RVFL component concatenates the direct input block $\boldsymbol{v}^{(\alpha)}$ and the randomized hidden-feature block $\boldsymbol{H}$ as $\boldsymbol{P} = [\boldsymbol{v}^{(\alpha)}, \boldsymbol{H}]$. The resulting forcing term drives the grey response, followed by CFD restoration. The red connections represent the direct input–output link, the blue connections represent the hidden-to-output connections whose coefficients form $\boldsymbol{\beta}$, and the black connections represent the fixed random input-to-hidden weights and biases w_j and b_j.

4. Optimization algorithm

The numerical algorithm for solving (3.23) is derived below. The objective contains two sources of difficulty: the ℓ_1/ℓ_2 norm-ratio regularizer $R(\boldsymbol{\beta})$ is non-smooth and non-convex, and the bounded exponential-squared loss is non-convex. Following the optimization structure of the non-convex ℓ_1/ℓ_2 exponential-loss ELM [36], the proposed solver uses ADMM variable splitting with a DCA update for the bounded-loss block. The hidden-output block $\boldsymbol{\beta}$ is updated using the analytically characterized

ℓ_1/ℓ_2 proximal mapping [38], while the direct-link coefficient γ is updated jointly with the grey coefficient and the auxiliary hidden-output vector. The DCA update majorizes only the exponential-squared-loss block. This distinction is essential: the norm-ratio proximal step is not a soft-thresholding DCA approximation.

4.1. Scaled equivalent formulation

Let $\boldsymbol{y}^{(\alpha)}$, $\boldsymbol{z}^{(\alpha)}$, $\boldsymbol{v}^{(\alpha)}$, and $\boldsymbol{H}$ be defined in (3.12)–(3.15). The vector $\boldsymbol{z}^{(\alpha)}$ enters the residual through the product $a\boldsymbol{z}^{(\alpha)}$. If the entries of $\boldsymbol{z}^{(\alpha)}$ have a much larger or much smaller scale than the direct and hidden RVFL features, the development coefficient a is estimated on an ill-conditioned numerical scale. To avoid this imbalance, introduce a small scaling safeguard $\epsilon_z > 0$. With $\mathbf{1} \in \mathbb{R}^{N-1}$ denoting the all-ones vector, define

$$\bar{z} = \frac{\mathbf{1}^\top \boldsymbol{z}^{(\alpha)}}{N-1}, \qquad s_z = \max\left\{\left[\frac{\|\boldsymbol{z}^{(\alpha)} - \bar{z}\mathbf{1}\|_2^2}{N-1}\right]^{1/2}, \epsilon_z\right\}. \tag{4.1}$$

The safeguard prevents division by a standard deviation close to zero when the background vector is close to constant; in the numerical experiments, ϵ_z is fixed at 10^{-8}. Define

$$\widetilde{\boldsymbol{z}}^{(\alpha)} = \frac{\boldsymbol{z}^{(\alpha)}}{s_z}, \qquad \widetilde{a} = a s_z.$$

Since $s_z > 0$,

$$a\boldsymbol{z}^{(\alpha)} = \frac{\widetilde{a}}{s_z}\boldsymbol{z}^{(\alpha)} = \widetilde{a}\widetilde{\boldsymbol{z}}^{(\alpha)}.$$

Thus, the transformation changes only the parameter scale and leaves the fitted residual unchanged. Substituting this identity into (3.23) gives the scaled but equivalent problem

$$\min_{\widetilde{a},\gamma\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^L} \Phi_\sigma\left(\boldsymbol{y}^{(\alpha)} + \widetilde{a}\widetilde{\boldsymbol{z}}^{(\alpha)} - \gamma\boldsymbol{v}^{(\alpha)} - \boldsymbol{H}\boldsymbol{\beta}\right) + \lambda R(\boldsymbol{\beta}). \tag{4.2}$$

After optimization, the development coefficient in the original scale is recovered by $\widehat{a} = \widehat{\widetilde{a}}/s_z$.

For notational simplicity, set $\boldsymbol{y} = \boldsymbol{y}^{(\alpha)}$, $\boldsymbol{z} = \widetilde{\boldsymbol{z}}^{(\alpha)}$, and $\boldsymbol{v} = \boldsymbol{v}^{(\alpha)}$. The scaled optimization problem can then be written as

$$\min_{\widetilde{a},\gamma\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^L} \Phi_\sigma\left(\boldsymbol{y} + \widetilde{a}\boldsymbol{z} - \gamma\boldsymbol{v} - \boldsymbol{H}\boldsymbol{\beta}\right) + \lambda R(\boldsymbol{\beta}). \tag{4.3}$$

4.2. ADMM splitting

In (4.3), the robust loss and the norm-ratio regularizer both depend on the hidden-output vector $\boldsymbol{\beta}$. The norm-ratio regularizer is easier to handle in a proximal form, whereas the robust loss is easier to approximate in the residual variable. Introduce an auxiliary vector $\boldsymbol{u} \in \mathbb{R}^L$ for the hidden-output coefficients inside the residual and enforce $\boldsymbol{\beta} = \boldsymbol{u}$. The direct coefficient γ requires no duplicate because it is absent from R. This yields

$$\begin{aligned} &\min_{\substack{\widetilde{a},\gamma\in\mathbb{R},\ \boldsymbol{u}\in\mathbb{R}^L \\ \boldsymbol{\beta}\in\mathbb{R}^L}} \Phi_\sigma\left(\boldsymbol{y} + \widetilde{a}\boldsymbol{z} - \gamma\boldsymbol{v} - \boldsymbol{H}\boldsymbol{u}\right) + \lambda R(\boldsymbol{\beta}), \\ &\text{s.t.} \qquad \boldsymbol{\beta} = \boldsymbol{u}. \end{aligned} \tag{4.4}$$

The equality constraint preserves the original feasible set because every feasible point of (4.3) corresponds to $(\widetilde{a}, \gamma, \boldsymbol{u}, \boldsymbol{\beta}) = (\widetilde{a}, \gamma, \boldsymbol{\beta}, \boldsymbol{\beta})$ in (4.4), and conversely every feasible point of (4.4) satisfies $\boldsymbol{u} = \boldsymbol{\beta}$.

For an ADMM penalty parameter $\eta > 0$ and multiplier $\boldsymbol{\mu} \in \mathbb{R}^L$, the augmented Lagrangian function is

$$\mathcal{L}_\eta(\widetilde{a}, \gamma, \boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\mu}) = \Phi_\sigma(\boldsymbol{y} + \widetilde{a}\boldsymbol{z} - \gamma\boldsymbol{v} - \boldsymbol{H}\boldsymbol{u}) + \lambda R(\boldsymbol{\beta}) + \langle \boldsymbol{\mu}, \boldsymbol{\beta} - \boldsymbol{u} \rangle + \frac{\eta}{2}\|\boldsymbol{\beta} - \boldsymbol{u}\|_2^2. \tag{4.5}$$

At iteration t, ADMM minimizes (4.5) with respect to one block while keeping the other block fixed, and then updates the multiplier. The two primal blocks are chosen as $\boldsymbol{\beta}$ and $(\widetilde{a}, \gamma, \boldsymbol{u})$.

The algorithm is initialized from the least-squares form of the discrete grey equation. Since the scaled residual is

$$\boldsymbol{r}(\widetilde{a}, \gamma, \boldsymbol{\beta}) = \boldsymbol{y} + \widetilde{a}\boldsymbol{z} - \gamma\boldsymbol{v} - \boldsymbol{H}\boldsymbol{\beta},$$

setting this residual approximately to zero gives

$$\boldsymbol{y} = -\widetilde{a}\boldsymbol{z} + \gamma\boldsymbol{v} + \boldsymbol{H}\boldsymbol{\beta}.$$

Define the design matrix

$$\boldsymbol{X} = \begin{bmatrix} -\boldsymbol{z} & \boldsymbol{v} & \boldsymbol{H} \end{bmatrix} \in \mathbb{R}^{(N-1)\times(L+2)}.$$

For a stabilizing constant $\varepsilon_0 > 0$, the ridge-stabilized least-squares estimate is

$$\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \varepsilon_0 \, \mathrm{Id}_{L+2}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y},$$

where Id_{L+2} is the $(L+2) \times (L+2)$ identity matrix. Set

$$\widetilde{a}^{(0)} = \theta_1^{(0)}, \qquad \gamma^{(0)} = \theta_2^{(0)}, \qquad \boldsymbol{\beta}^{(0)} = \left(\theta_3^{(0)}, \ldots, \theta_{L+2}^{(0)}\right)^\top, \qquad \boldsymbol{u}^{(0)} = \boldsymbol{\beta}^{(0)}, \qquad \boldsymbol{\mu}^{(0)} = \boldsymbol{0}.$$

The extension in (3.21) permits $\boldsymbol{\beta}^{(0)} = \boldsymbol{0}$; hence no direction-dependent perturbation is introduced when the ridge initializer is zero.

Given $(\widetilde{a}^{(t)}, \gamma^{(t)}, \boldsymbol{u}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\mu}^{(t)})$, the terms in (4.5) that depend on $\boldsymbol{\beta}$ are

$$\lambda R(\boldsymbol{\beta}) + \left\langle \boldsymbol{\mu}^{(t)}, \boldsymbol{\beta} - \boldsymbol{u}^{(t)} \right\rangle + \frac{\eta}{2} \left\| \boldsymbol{\beta} - \boldsymbol{u}^{(t)} \right\|_2^2.$$

The terms independent of $\boldsymbol{\beta}$ can be dropped. Completing the square gives

$$\left\langle \boldsymbol{\mu}^{(t)}, \boldsymbol{\beta} - \boldsymbol{u}^{(t)} \right\rangle + \frac{\eta}{2} \left\| \boldsymbol{\beta} - \boldsymbol{u}^{(t)} \right\|_2^2 = \frac{\eta}{2} \left\| \boldsymbol{\beta} - \boldsymbol{u}^{(t)} + \frac{\boldsymbol{\mu}^{(t)}}{\eta} \right\|_2^2 + \text{const.}$$

The $\boldsymbol{\beta}$-subproblem is

$$\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^L} \lambda R(\boldsymbol{\beta}) + \frac{\eta}{2} \left\| \boldsymbol{\beta} - \boldsymbol{u}^{(t)} + \frac{\boldsymbol{\mu}^{(t)}}{\eta} \right\|_2^2. \tag{4.6}$$

Set $\boldsymbol{q}^{(t)} = \boldsymbol{u}^{(t)} - \boldsymbol{\mu}^{(t)}/\eta$ and $\rho = \eta/\lambda$. After division by the positive constant λ, (4.6) is equivalent to the ratio-penalized proximal problem

$$\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^L} R(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{q}^{(t)}\|_2^2. \tag{4.7}$$

4.3. ℓ_1/ℓ_2 proximal update

The minimization in (4.7) is the proximal problem studied by Tao [38] and subsequently used in the non-convex ELM of Wang and Wang [36]. Its solution is not the ordinary ℓ_1 soft-thresholding operator because the denominator $\|\boldsymbol{x}\|_2$ couples all coordinates. The required characterization is restated below in the notation of the present model.

Proposition 4.1 (Characterization of the ℓ_1/ℓ_2 proximal point)**.** For $\rho > 0$ and $\boldsymbol{q} \in \mathbb{R}^L$, define

$$\mathrm{Prox}_R^{\rho}(\boldsymbol{q}) = \arg\min_{\boldsymbol{x}\in\mathbb{R}^L} \left\{ R(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{q}\|_2^2 \right\}.$$

If $\boldsymbol{q} = \boldsymbol{0}$, then $\mathrm{Prox}_R^{\rho}(\boldsymbol{0}) = \{\boldsymbol{0}\}$. Suppose $\boldsymbol{q} \neq \boldsymbol{0}$, and let ϖ be a permutation such that

$$|q_{\varpi(1)}| \geq |q_{\varpi(2)}| \geq \cdots \geq |q_{\varpi(L)}|, \qquad |q_{\varpi(L+1)}| := 0. \tag{4.8}$$

If $0 < \rho \leq |q_{\varpi(1)}|^{-2}$, one proximal point is the one-sparse vector

$$\bar{x}_{\varpi(i)} = \begin{cases} q_{\varpi(j)}, & i = j, \\ 0, & i \neq j, \end{cases} \tag{4.9}$$

where $j \in \{1, \ldots, s_{\max}\}$ and $s_{\max}$ is the multiplicity of the largest magnitude in (4.8). If $\rho > |q_{\varpi(1)}|^{-2}$, there is a support size $d \in \{1, \ldots, L\}$ and positive scalars $p = \|\bar{\boldsymbol{x}}\|_1$ and $r = \|\bar{\boldsymbol{x}}\|_2$ such that

$$\frac{p^2}{r^3} - \rho p + \rho Q_{1,d} - \frac{d}{r} = 0, \tag{4.10a}$$

$$r^3 - Q_{2,d} r + \frac{Q_{1,d} - p}{\rho} = 0, \tag{4.10b}$$

where

$$Q_{1,d} = \sum_{i=1}^{d} |q_{\varpi(i)}|, \qquad Q_{2,d} = \sum_{i=1}^{d} q_{\varpi(i)}^2. \tag{4.11}$$

The support is compatible with (p, r) when

$$|q_{\varpi(d)}| > \frac{1}{\rho r}, \qquad |q_{\varpi(d+1)}| \leq \frac{1}{\rho r}, \qquad \rho > \frac{p}{r^3}. \tag{4.12}$$

For such a support, a proximal point is

$$\bar{x}_{\varpi(i)} = \begin{cases} \mathrm{sign}(q_{\varpi(i)}) \dfrac{\rho|q_{\varpi(i)}| - 1/r}{\rho - p/r^3}, & 1 \leq i \leq d, \\ 0, & d < i \leq L. \end{cases} \tag{4.13}$$

The origin of the two scalar Eqs (4.10a) and (4.10b) can be explained through the first-order optimality (Karush–Kuhn–Tucker) condition of the proximal objective. Since (4.7) is an unconstrained non-smooth problem, the zero vector must belong to the subdifferential of its objective at a minimizer

$\bar{\boldsymbol{x}} \neq \mathbf{0}$. With $p = \|\bar{\boldsymbol{x}}\|_1$ and $r = \|\bar{\boldsymbol{x}}\|_2$, the gradient identity $\nabla(1/\|\boldsymbol{x}\|_2) = -\boldsymbol{x}/\|\boldsymbol{x}\|_2^3$ and the sign subdifferential $\boldsymbol{\zeta} \in \partial\|\bar{\boldsymbol{x}}\|_1$ turn this condition into

$$\mathbf{0} \in \frac{\boldsymbol{\zeta}}{r} - \frac{p}{r^3}\bar{\boldsymbol{x}} + \rho(\bar{\boldsymbol{x}} - \boldsymbol{q}). \tag{4.14}$$

The scalars p and r couple all coordinates, which is why the solution is not an ordinary soft-thresholding step. The quadratic term is minimized when the signs and descending magnitude order of $\bar{\boldsymbol{x}}$ agree with those of $\boldsymbol{q}$. For an active coordinate $\varpi(i)$, $\zeta_{\varpi(i)} = \operatorname{sign}(q_{\varpi(i)})$, and (4.14) can be solved coordinate-wise for $|\bar{x}_{\varpi(i)}|$:

$$\left(\rho - \frac{p}{r^3}\right)|\bar{x}_{\varpi(i)}| = \rho|q_{\varpi(i)}| - \frac{1}{r}.$$

This identity gives (4.13). Equations (4.10a) and (4.10b) then follow by requiring the coordinate-wise solution to be consistent with the definitions of the two auxiliary scalars: summing $|\bar{x}_{\varpi(i)}|$ over the d active coordinates and equating the sum to p yields (4.10a), whereas summing $\bar{x}^2_{\varpi(i)}$ over the same coordinates and equating the sum to r^2 yields (4.10b). For an inactive coordinate, the subgradient condition $|\zeta_{\varpi(i)}| \leq 1$ gives $|q_{\varpi(i)}| \leq 1/(\rho r)$, whereas positivity of an active numerator gives the strict reverse inequality. These two facts produce (4.12). The remaining one-sparse case and the threshold $|q_{\varpi(1)}|^{-2}$ follow from the global proximal characterization in [38].

For a fixed trial support size d, the two scalar equations in (4.10) can be solved by alternately selecting the smaller root for p and the largest positive root for r. Starting from $r^{(0)} > 0$, define

$$\Delta^{(2)}_{j+1} = \rho^2 - \frac{4}{(r^{(j)})^3}\left(\rho Q_{1,d} - \frac{d}{r^{(j)}}\right), \tag{4.15a}$$

$$p^{(j+1)} = \frac{(r^{(j)})^3}{2}\left(\rho - \sqrt{\Delta^{(2)}_{j+1}}\right), \tag{4.15b}$$

$$\upsilon_d = \sqrt{\frac{Q_{2,d}}{3}}, \qquad \operatorname{ang}^{(j+1)} = \arccos\left(\frac{Q_{1,d} - p^{(j+1)}}{2\rho\upsilon_d^3}\right), \tag{4.15c}$$

$$r^{(j+1)} = 2\upsilon_d \cos\left(\frac{\pi}{3} - \frac{\operatorname{ang}^{(j+1)}}{3}\right). \tag{4.15d}$$

The smaller quadratic root in (4.15b) and the largest positive cubic root in (4.15d) are selected because the ratio p/r is reduced by a smaller p and a larger r. In finite-precision arithmetic, a trial support is rejected if the discriminant is materially negative or the argument of arccos lies outside $[-1, 1]$ beyond a roundoff tolerance. Otherwise, iteration stops when

$$\frac{|r^{(j+1)} - r^{(j)}|}{\max\{1, |r^{(j)}|\}} \leq \varepsilon_R, \tag{4.16}$$

or after $J_{\max}$ scalar iterations. The support size is then located by the bisection logic in Algorithm 1. This is the practical support-search procedure proposed in [38]; as explicitly noted there, the bisection implementation is heuristic even though the proximal-point characterization in Proposition 4.1 is analytical.

Algorithm 1. Practical ℓ_1/ℓ_2 proximal solver.

Input: $\boldsymbol{q} \in \mathbb{R}^L$, $\rho > 0$, scalar-iteration limit $J_{\max}$, tolerance $\varepsilon_R > 0$, and initial radius $r_{\text{init}} > 0$.

Output: An approximate ℓ_1/ℓ_2 proximal point $\bar{\boldsymbol{x}}$.

if $\boldsymbol{q} = \mathbf{0}$ **then**
  return $\bar{\boldsymbol{x}} = \mathbf{0}$;
Sort $\boldsymbol{q}$ according to (4.8) and set $|q_{\varpi(L+1)}| = 0$;
if $0 < \rho \le |q_{\varpi(1)}|^{-2}$ **then**
  Construct $\bar{\boldsymbol{x}}$ by (4.9) and **return** $\bar{\boldsymbol{x}}$;
Set $d_{\text{lo}} = 1$, $d_{\text{hi}} = L$, and initialize an empty candidate set C;
while $d_{\text{hi}} - d_{\text{lo}} > 1$ **do**
  Set $d = \lfloor (d_{\text{lo}} + d_{\text{hi}})/2 \rfloor$ and compute $Q_{1,d}, Q_{2,d}$ by (4.11);
  Starting from $r^{(0)} = r_{\text{init}}$, iterate (4.15) until (4.16), $J_{\max}$, or an invalid real-root condition is reached;
  if *the scalar iteration gives a valid pair* $(\widehat{p}, \widehat{r})$ **then**
    Compute $\text{gap}_d = (|q_{\varpi(d)}| - 1/(\rho\widehat{r}))(1 - \widehat{p}/(\rho\widehat{r}^3))$ and define gap_{d+1} analogously;
    if $\text{gap}_d > 0$ *and* $\text{gap}_{d+1} \le 0$ **then**
      Construct a candidate by (4.13), add it to C, and **break**;
    else
      if $\text{gap}_d \le 0$ **then**
        Set $d_{\text{hi}} = d$;
      else
        Set $d_{\text{lo}} = d$;
  else
    Using the last positive radius $\widehat{r}$, set $d_{\text{lo}} = d$ if $|q_{\varpi(d+1)}| > 1/(\rho\widehat{r})$; otherwise set $d_{\text{hi}} = d$;
Test the two endpoint supports in $\{d_{\text{lo}}, d_{\text{hi}}\}$ by (4.12)–(4.15) and add every valid candidate to C;
if $C = \varnothing$ **then**
  Test all remaining support sizes $d \in \{1, \ldots, L\}$;
  Report numerical failure and stop if no valid real-root candidate is found;
return *the candidate in* C *with the smallest value of* $R(\boldsymbol{x}) + (\rho/2)\|\boldsymbol{x} - \boldsymbol{q}\|_2^2$.

Remark 4.2. Proposition 4.1 characterizes a global minimizer of the $\boldsymbol{\beta}$-subproblem, but it does not by itself establish global convergence of the complete CFSRGRVFL iteration sequence. Tao's convergence theorem applies to a specific least-squares model and splitting scheme [38], whereas the present model contains an additional bounded non-convex loss updated by DCA. Accordingly, the full procedure is treated here as a numerically evaluated algorithm; no unproved global-convergence claim is made.

After the sparse hidden-output update, the variables $\widetilde{a}$, γ, and $\boldsymbol{u}$ are updated jointly. Keeping only the terms in (4.5) that depend on $(\widetilde{a}, \gamma, \boldsymbol{u})$ gives

$$\Phi_\sigma\left(\boldsymbol{y}+\widetilde{a}\boldsymbol{z}-\gamma\boldsymbol{v}-\boldsymbol{H}\boldsymbol{u}\right)-\left\langle\boldsymbol{\mu}^{(t)},\boldsymbol{u}\right\rangle+\frac{\eta}{2}\left\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{u}\right\|_2^2.$$

Completing the square with respect to $\boldsymbol{u}$ gives

$$-\left\langle\boldsymbol{\mu}^{(t)},\boldsymbol{u}\right\rangle+\frac{\eta}{2}\left\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{u}\right\|_2^2=\frac{\eta}{2}\left\|\boldsymbol{u}-\boldsymbol{c}^{(t+1)}\right\|_2^2+\text{const},$$

where

$$\boldsymbol{c}^{(t+1)}=\boldsymbol{\beta}^{(t+1)}+\frac{\boldsymbol{\mu}^{(t)}}{\eta}.$$

Thus, the exact $(\widetilde{a},\gamma,\boldsymbol{u})$ subproblem is

$$\min_{\widetilde{a},\gamma\in\mathbb{R},\ \boldsymbol{u}\in\mathbb{R}^L}\Phi_\sigma\left(\boldsymbol{y}+\widetilde{a}\boldsymbol{z}-\gamma\boldsymbol{v}-\boldsymbol{H}\boldsymbol{u}\right)+\frac{\eta}{2}\left\|\boldsymbol{u}-\boldsymbol{c}^{(t+1)}\right\|_2^2.$$

To express this problem compactly, define

$$\boldsymbol{\chi}=(\widetilde{a},\gamma,\boldsymbol{u}^\top)^\top,\qquad \boldsymbol{G}=[\boldsymbol{z}\quad -\boldsymbol{v}\quad -\boldsymbol{H}],\qquad \boldsymbol{D}=[\boldsymbol{0}_{L\times 2}\quad \mathrm{Id}_L].$$

Then, $\boldsymbol{\chi}\in\mathbb{R}^{L+2}$, $\boldsymbol{G}\in\mathbb{R}^{(N-1)\times(L+2)}$, $\boldsymbol{D}\in\mathbb{R}^{L\times(L+2)}$, and

$$\boldsymbol{y}+\widetilde{a}\boldsymbol{z}-\gamma\boldsymbol{v}-\boldsymbol{H}\boldsymbol{u}=\boldsymbol{y}+\boldsymbol{G}\boldsymbol{\chi},\qquad \boldsymbol{u}=\boldsymbol{D}\boldsymbol{\chi}.$$

The scalar robust loss admits the difference-of-convex (DC) decomposition

$$\phi_\sigma(r)=\frac{1}{2}r^2-\psi_\sigma(r),$$

where

$$\psi_\sigma(r)=\frac{1}{2}r^2-\frac{\sigma^2}{2}\left(1-\exp\left(-\frac{r^2}{\sigma^2}\right)\right).$$

The derivative of the convex component ψ_σ is

$$\psi'_\sigma(r)=r\left(1-\exp\left(-\frac{r^2}{\sigma^2}\right)\right).$$

Indeed,

$$\psi''_\sigma(r)=1+\left(\frac{2r^2}{\sigma^2}-1\right)\exp\left(-\frac{r^2}{\sigma^2}\right)\geq 0.$$

Writing $\chi_0=r^2/\sigma^2\geq 0$, denote the righthand side by $\Gamma(\chi_0)=1+(2\chi_0-1)e^{-\chi_0}$; because $\Gamma(0)=0$, $\Gamma'(\chi_0)=(3-2\chi_0)e^{-\chi_0}$, and $\Gamma(\chi_0)\to 1$ as $\chi_0\to\infty$, it is nonnegative on $[0,\infty)$. Hence, ψ_σ is convex, as required by the DC decomposition. Let $\odot$ denote component-wise multiplication. Consider outer iteration t. Define the current residual by

$$\boldsymbol{r}^{(t)}=\boldsymbol{y}+\widetilde{a}^{(t)}\boldsymbol{z}-\gamma^{(t)}\boldsymbol{v}-\boldsymbol{H}\boldsymbol{u}^{(t)}.$$

The vector of first derivatives of ψ_σ evaluated at $\boldsymbol{r}^{(t)}$ is

$$\boldsymbol{s}^{(t)}=\boldsymbol{r}^{(t)}\odot\left(\boldsymbol{1}-\exp\left(-\frac{\boldsymbol{r}^{(t)}\odot\boldsymbol{r}^{(t)}}{\sigma^2}\right)\right).$$

Using $\Phi_\sigma(\boldsymbol{r}) = \frac{1}{2}\|\boldsymbol{r}\|_2^2 - \sum_i \psi_\sigma(r_i)$, convexity of ψ_σ gives

$$-\sum_i \psi_\sigma(r_i) \le -\sum_i \psi_\sigma(r_i^{(t)}) - \left\langle \boldsymbol{s}^{(t)}, \boldsymbol{r} - \boldsymbol{r}^{(t)} \right\rangle. \tag{4.17}$$

Equality holds at $\boldsymbol{r} = \boldsymbol{r}^{(t)}$. Thus, (4.17) is a global affine majorizer of the concave component rather than a merely formal first-order approximation. After removing constants independent of $\boldsymbol{\chi}$, the resulting convex DCA subproblem is

$$\boldsymbol{\chi}^{(t+1)} = \operatorname*{arg\,min}_{\boldsymbol{\chi}\in\mathbb{R}^{L+2}} \frac{1}{2}\|\boldsymbol{y} + \boldsymbol{G}\boldsymbol{\chi} - \boldsymbol{s}^{(t)}\|_2^2 + \frac{\eta}{2}\|\boldsymbol{D}\boldsymbol{\chi} - \boldsymbol{c}^{(t+1)}\|_2^2. \tag{4.18}$$

The quadratic part satisfies

$$\frac{1}{2}\|\boldsymbol{y} + \boldsymbol{G}\boldsymbol{\chi}\|_2^2 - \left\langle \boldsymbol{s}^{(t)}, \boldsymbol{y} + \boldsymbol{G}\boldsymbol{\chi} \right\rangle = \frac{1}{2}\|\boldsymbol{y} + \boldsymbol{G}\boldsymbol{\chi} - \boldsymbol{s}^{(t)}\|_2^2 + \text{const.}$$

Define

$$\boldsymbol{A}_\eta = \boldsymbol{G}^\top\boldsymbol{G} + \eta\boldsymbol{D}^\top\boldsymbol{D}, \qquad \boldsymbol{b}^{(t+1)} = \boldsymbol{G}^\top(\boldsymbol{s}^{(t)} - \boldsymbol{y}) + \eta\boldsymbol{D}^\top\boldsymbol{c}^{(t+1)}.$$

The normal equations for (4.18) are $\boldsymbol{A}_\eta\boldsymbol{\chi} = \boldsymbol{b}^{(t+1)}$. Their solvability follows from the least-squares form of (4.18). Moreover, for any $\boldsymbol{d} = (d_a, d_\gamma, \boldsymbol{d}_u^\top)^\top$,

$$\boldsymbol{d}^\top\boldsymbol{A}_\eta\boldsymbol{d} = \|d_a\boldsymbol{z} - d_\gamma\boldsymbol{v} - \boldsymbol{H}\boldsymbol{d}_u\|_2^2 + \eta\|\boldsymbol{d}_u\|_2^2.$$

Hence, $\boldsymbol{A}_\eta$ is positive definite whenever $\operatorname{rank}[\boldsymbol{z}\ \boldsymbol{v}] = 2$: equality can then hold only if $\boldsymbol{d}_u = \boldsymbol{0}$ and $d_a\boldsymbol{z} - d_\gamma\boldsymbol{v} = \boldsymbol{0}$, which implies $d_a = d_\gamma = 0$. In this identifiable case, the DCA subproblem has a unique minimizer. If $\operatorname{rank}[\boldsymbol{z}\ \boldsymbol{v}] < 2$, the coefficients $\widetilde{a}$ and γ are not separately identifiable from this sample; the Moore–Penrose inverse selects the minimum-norm minimizer without changing the quadratic subproblem. The update is therefore

$$\boldsymbol{\chi}^{(t+1)} = \begin{cases} \boldsymbol{A}_\eta^{-1}\boldsymbol{b}^{(t+1)}, & \operatorname{rank}[\boldsymbol{z}\ \boldsymbol{v}] = 2, \\ \boldsymbol{A}_\eta^{\dagger}\boldsymbol{b}^{(t+1)}, & \operatorname{rank}[\boldsymbol{z}\ \boldsymbol{v}] < 2. \end{cases} \tag{4.19}$$

The first component of $\boldsymbol{\chi}^{(t+1)}$ is $\widetilde{a}^{(t+1)}$, the second is $\gamma^{(t+1)}$, and the remaining L components form $\boldsymbol{u}^{(t+1)}$.

4.4. Multiplier update and stopping rule

The multiplier update is

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \eta\left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{u}^{(t+1)}\right).$$

The primal residual measures violation of the splitting constraint $\boldsymbol{\beta} = \boldsymbol{u}$, whereas the dual residual measures the change in the auxiliary block. They are defined as

$$\mathrm{res}_{\mathrm{pri}}^{(t+1)} = \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{u}^{(t+1)}\|_2, \qquad \mathrm{res}_{\mathrm{dual}}^{(t+1)} = \eta\|\boldsymbol{u}^{(t+1)} - \boldsymbol{u}^{(t)}\|_2.$$

Let $K_{\min}, K_{\max} \in \mathbb{Z}_{\ge 0}$ satisfy $0 \le K_{\min} \le K_{\max}$, and let $\varepsilon_{\mathrm{ADMM}} > 0$. The outer iteration is terminated after at least $K_{\min}$ iterations once both residuals are no larger than $\varepsilon_{\mathrm{ADMM}}$; otherwise, it is terminated after $K_{\max}$ iterations.

After termination, set $\widehat{\widetilde{a}}$, $\widehat{\gamma}$, and $\widehat{\boldsymbol{\beta}}$ equal to the final iterates and recover $\widehat{a} = \widehat{\widetilde{a}}/s_z$. The fitted accumulated sequence is generated by (3.10), equivalently by the recursion (3.11), with $(a, \gamma, \boldsymbol{\beta})$ replaced by $(\widehat{a}, \widehat{\gamma}, \widehat{\boldsymbol{\beta}})$. The fitted original sequence is then obtained from (3.3).

Algorithm 2 summarizes the resulting procedure.

Algorithm 2. ADMM-based algorithm with a DCA update for the CFSRGRVFL model.

Input: Original sequence $\boldsymbol{T}^{(0)} \in \mathbb{R}^N_{\geq 0}$; hyperparameters $\alpha \in (0, 1]$, $L \in \mathbb{N}$, $\lambda > 0$, $\sigma > 0$, $\eta > 0$; numerical constants $\epsilon_z, \varepsilon_0, r_{\text{init}} > 0$; iteration limits $K_{\min}, K_{\max} \in \mathbb{Z}_{\geq 0}$ and $J_{\max} \in \mathbb{N}$ with $K_{\min} \leq K_{\max}$; tolerances $\varepsilon_{\text{ADMM}}, \varepsilon_R > 0$.

Output: Estimated parameters $\widehat{a}$, $\widehat{\gamma}$, and $\widehat{\boldsymbol{\beta}}$, and fitted sequence $\widehat{\boldsymbol{T}}^{(0)}$.

Construct $\boldsymbol{T}^{(\alpha)}$, $\boldsymbol{y}^{(\alpha)}$, $\boldsymbol{z}^{(\alpha)}$, the direct-input vector $\boldsymbol{v}^{(\alpha)}$, and the hidden-feature matrix $\boldsymbol{H}$ as in Section 3.2;

Compute s_z and $\widetilde{\boldsymbol{z}}^{(\alpha)}$ by (4.1), and set $\boldsymbol{y} = \boldsymbol{y}^{(\alpha)}$, $\boldsymbol{z} = \widetilde{\boldsymbol{z}}^{(\alpha)}$, and $\boldsymbol{v} = \boldsymbol{v}^{(\alpha)}$;

Initialize $\widetilde{a}^{(0)}$, $\gamma^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{u}^{(0)}$, and $\boldsymbol{\mu}^{(0)}$ by ridge-stabilized least squares;

Form $\boldsymbol{G} = [\boldsymbol{z} \ -\boldsymbol{v} \ -\boldsymbol{H}]$ and $\boldsymbol{D} = [\boldsymbol{0}_{L\times 2} \ \mathrm{Id}_L]$;

for $t = 0$ **to** $K_{\max} - 1$ **do**

  Set $\boldsymbol{q}^{(t)} = \boldsymbol{u}^{(t)} - \boldsymbol{\mu}^{(t)}/\eta$ and $\rho = \eta/\lambda$;

  Compute $\boldsymbol{\beta}^{(t+1)}$ by Algorithm 1 using $(\boldsymbol{q}^{(t)}, \rho, J_{\max}, \varepsilon_R, r_{\text{init}})$;

  Compute $\boldsymbol{s}^{(t)}$ from the current residual $\boldsymbol{r}^{(t)} = \boldsymbol{y} + \widetilde{a}^{(t)}\boldsymbol{z} - \gamma^{(t)}\boldsymbol{v} - \boldsymbol{H}\boldsymbol{u}^{(t)}$;

  Set $\boldsymbol{c}^{(t+1)} = \boldsymbol{\beta}^{(t+1)} + \boldsymbol{\mu}^{(t)}/\eta$;

  Compute $\boldsymbol{\chi}^{(t+1)} = (\widetilde{a}^{(t+1)}, \gamma^{(t+1)}, (\boldsymbol{u}^{(t+1)})^\top)^\top$ by (4.19);

  Update $\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} + \eta(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{u}^{(t+1)})$;

  Compute $\mathrm{res}_{\text{pri}}^{(t+1)}$ and $\mathrm{res}_{\text{dual}}^{(t+1)}$;

  if $t + 1 \geq K_{\min}$ *and* $\mathrm{res}_{\text{pri}}^{(t+1)} \leq \varepsilon_{\text{ADMM}}$ *and* $\mathrm{res}_{\text{dual}}^{(t+1)} \leq \varepsilon_{\text{ADMM}}$ **then**

    break;

Set $\widehat{\widetilde{a}} = \widetilde{a}^{(t+1)}$, $\widehat{\gamma} = \gamma^{(t+1)}$, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t+1)}$, and $\widehat{a} = \widehat{\widetilde{a}}/s_z$;

Generate $\widehat{\boldsymbol{T}}^{(\alpha)}$ by the recursion (3.11), and restore $\widehat{\boldsymbol{T}}^{(0)}$ by (3.3).

For an integer forecast horizon $N_{\text{fc}} \geq 1$, the time index is extended to $k = N + 1, N + 2, \ldots, N + N_{\text{fc}}$ while the normalization length N_s is kept fixed as specified in Section 3.2. The CFA time inputs are constructed on the extended index set and standardized with the training values (μ_v, s_v). Each future RVFL forcing value is $\widehat{\gamma}\widetilde{v}_\tau^{(\alpha)}(k) + \sum_{j=1}^{L} \widehat{\beta}_j h_j(k)$. The exponential recursion (3.11) produces the accumulated forecasts, and (3.3) produces the corresponding original-scale forecasts.

5. Bayesian hyperparameter optimization

The hyperparameters are optimized in an outer loop around the ADMM-based coefficient-estimation problem. For a candidate hyperparameter vector and a prescribed random seed, Algorithm 2 estimates only the grey development coefficient, the RVFL direct-link coefficient, and the hidden-output weights. The outer problem evaluates the restored modeling-sample fit and updates the

hyperparameter vector. This separation yields a nested optimization scheme with hyperparameter optimization in the outer loop and coefficient estimation in the inner loop.

5.1. Coupled hyperparameters and nested optimization

The optimization protocol is designed for short modeling sequences in grey forecasting. Ma et al. [29] selected the hidden-layer size through validation-based grid search, whereas Gong and Zong [35] selected grey RVFL hyperparameters using errors computed from the modeling sequence. In the present setting, further partitioning an already short modeling sequence would leave fewer than $N_{\mathrm{tr}} - 1$ grey balance equations for coefficient estimation. More importantly, every truncation reconstructs the α-dependent accumulated response, background vector, standardized direct feature, and hidden-feature matrix. Such truncated problems do not preserve the design used by the final model. The complete modeling sequence is therefore retained to evaluate the outer objective and to fit the final model, while the forecasting sequence is excluded from both operations.

Let $\boldsymbol{T}_{\mathrm{tr}}^{(0)} \in \mathbb{R}_{\geq 0}^{N_{\mathrm{tr}}}$ be the modeling sequence, where $N_{\mathrm{tr}} \geq 2$. Define the hyperparameter vector as

$$\boldsymbol{\xi} = (\alpha, L, \lambda, \kappa_\sigma, \eta, c_w, c_b).$$

The CFA order α controls temporal weighting, L is the number of hidden nodes, and c_w and c_b are the scaling factors for the random hidden weights and biases. The regularization parameter λ controls the strength of hidden-feature selection. The dimensionless multiplier $\kappa_\sigma > 0$ determines the scale parameter of the bounded loss relative to the transformed-response scale, and $\eta > 0$ is the ADMM penalty parameter.

To normalize the loss-scale parameter across CFA orders and data scales, define

$$\bar{y}_{\mathrm{tr}}^{(\alpha)} = \frac{1}{N_{\mathrm{tr}} - 1} \sum_{k=2}^{N_{\mathrm{tr}}} \frac{T_{\mathrm{tr}}^{(0)}(k)}{k^{1-\alpha}},$$

and, for a fixed safeguard $\varepsilon_\sigma > 0$,

$$\begin{aligned} s_y(\alpha) &= \max\left\{\left[\frac{1}{N_{\mathrm{tr}} - 1} \sum_{k=2}^{N_{\mathrm{tr}}} \left(\frac{T_{\mathrm{tr}}^{(0)}(k)}{k^{1-\alpha}} - \bar{y}_{\mathrm{tr}}^{(\alpha)}\right)^2\right]^{1/2}, \varepsilon_\sigma\right\}, \\ \sigma(\alpha, \kappa_\sigma) &= \kappa_\sigma s_y(\alpha). \end{aligned} \tag{5.1}$$

Thus, κ_σ is the outer optimization variable, whereas the bounded loss in (3.23) uses the induced scale parameter $\sigma(\alpha, \kappa_\sigma) > 0$.

Let $\mathcal{N}_L \subset \mathbb{N}$ be a nonempty finite set of candidate hidden-layer sizes. Given $0 < \alpha_{\min} \leq 1$ and bounds satisfying

$$\begin{gathered} 0 < \lambda_{\min} \leq \lambda_{\max}, \qquad 0 < \kappa_{\sigma,\min} \leq \kappa_{\sigma,\max}, \qquad 0 < \eta_{\min} \leq \eta_{\max}, \\ 0 < c_{w,\min} \leq c_{w,\max}, \qquad 0 < c_{b,\min} \leq c_{b,\max}, \end{gathered}$$

define the mixed discrete–continuous search space

$$\Omega = [\alpha_{\min}, 1] \times \mathcal{N}_L \times [\lambda_{\min}, \lambda_{\max}] \times [\kappa_{\sigma,\min}, \kappa_{\sigma,\max}] \times [\eta_{\min}, \eta_{\max}] \times [c_{w,\min}, c_{w,\max}] \times [c_{b,\min}, c_{b,\max}].$$

The hyperparameters are not estimated jointly with $(a, \gamma, \boldsymbol{\beta})$ in (3.23). Instead, each pair $(\boldsymbol{\xi}, s_m)$ specifies one inner coefficient-estimation problem and one corresponding random-feature realization.

5.2. Objective averaged over fixed random seeds

Let $\mathcal{S} = \{s_1, \ldots, s_M\}$, with $M \geq 1$, be a finite set of random seeds prescribed before hyperparameter optimization and held fixed throughout the search. For a candidate $\boldsymbol{\xi}$ and seed s_m, construct the random features, set $\sigma = \sigma(\alpha, \kappa_\sigma)$ by (5.1), solve the inner problem, and restore the fitted original-scale sequence $\widehat{\boldsymbol{T}}^{(0)}_{\mathrm{tr},m}(\boldsymbol{\xi})$ using (3.3). Because $\widehat{T}^{(0)}_{\mathrm{tr},m}(1;\boldsymbol{\xi}) = T^{(0)}_{\mathrm{tr}}(1)$ is imposed by the time-response initial condition, the first observation has identically zero fitting error and is excluded from the outer objective. Define the original-scale modeling-sample error under seed s_m as

$$E_m(\boldsymbol{\xi}) = \frac{100}{N_{\mathrm{tr}} - 1} \sum_{k=2}^{N_{\mathrm{tr}}} \frac{\left| T^{(0)}_{\mathrm{tr}}(k) - \widehat{T}^{(0)}_{\mathrm{tr},m}(k;\boldsymbol{\xi}) \right|}{\max\left\{ \left| T^{(0)}_{\mathrm{tr}}(k) \right|, \varepsilon_{\mathrm{MAPE}} \right\}}.$$

Here $\varepsilon_{\mathrm{MAPE}} > 0$ prevents division by zero. The outer objective function is

$$\mathcal{J}(\boldsymbol{\xi}) = \frac{1}{M} \sum_{m=1}^{M} E_m(\boldsymbol{\xi}).$$

The same set of random seeds is used for every candidate. With deterministic inner solves, this construction makes repeated evaluations of $\mathcal{J}$ reproducible. Averaging over $\mathcal{S}$ reduces the dependence of the outer objective on a single random-feature realization and supports consistent comparisons among candidates. This average is part of the hyperparameter optimization objective and does not define a forecast ensemble. In contrast to the Top-k objective of Gong and Zong [35], E_m retains every noninitial fitted observation; the influence of extreme residuals on coefficient estimation is handled inside the model by the bounded exponential-squared loss.

The objective is finite at $\boldsymbol{\xi} \in \Omega$ whenever all M inner runs return finite fitted values. If a global minimum is attained, a global solution satisfies

$$\boldsymbol{\xi}^\star \in \arg\min_{\boldsymbol{\xi} \in \Omega} \mathcal{J}(\boldsymbol{\xi}).$$

The objective uses in-sample fitting errors from the complete modeling sequence and is not an independent estimator of generalization error. Forecasting performance is assessed only on the subsequent observations excluded from hyperparameter optimization and coefficient estimation.

5.3. Bayesian optimization and final model fitting

The outer problem contains both continuous and discrete variables, and one evaluation of $\mathcal{J}$ requires M complete ADMM–DCA runs. Analytic derivatives of this computationally expensive mixed-variable black-box objective are unavailable. Bayesian optimization [39] is therefore used to select evaluation points sequentially under a finite budget. After t evaluations, let $\mathcal{D}_t = \{(\boldsymbol{\xi}^{(i)}, \mathcal{J}(\boldsymbol{\xi}^{(i)}))\}_{i=1}^{t}$. A probabilistic surrogate model is fitted to $\mathcal{D}_t$ to approximate $\mathcal{J}(\boldsymbol{\xi})$. The expected-improvement acquisition function is

$$\mathrm{EI}_t(\boldsymbol{\xi}) = \mathbb{E}_t\left[\left(\mathcal{J}^{(t)}_{\min} - \widetilde{\mathcal{J}}_t(\boldsymbol{\xi}) \right)_+ \right].$$

The variable $\widetilde{\mathcal{J}}_t(\boldsymbol{\xi})$ denotes the surrogate random variable, $(s)_+ = \max\{s, 0\}$, and $\mathcal{J}^{(t)}_{\min} = \min_{1 \leq i \leq t} \mathcal{J}(\boldsymbol{\xi}^{(i)})$. The next candidate is selected as a maximizer of the acquisition function, $\boldsymbol{\xi}^{(t+1)} \in$

$\arg\max_{\xi\in\Omega} \mathrm{EI}_t(\boldsymbol{\xi})$. The objective value $\mathcal{J}(\boldsymbol{\xi}^{(t+1)})$ is then evaluated using the complete modeling sequence and the fixed set of random seeds, after which the surrogate model is updated. Under an evaluation budget N_{BO}, the incumbent is the evaluated hyperparameter vector with the smallest observed objective value. It is not asserted to be a global minimizer.

Algorithm 3 summarizes the Bayesian optimization procedure. A seed $s_{\mathrm{ref}} \in \mathcal{S}$ for final model fitting is prescribed before optimization and is not treated as an optimization variable. After the incumbent hyperparameter vector $\widehat{\boldsymbol{\xi}}$ is obtained, the final CFSRGRVFL model is fitted on the complete modeling sequence using $\widehat{\boldsymbol{\xi}}$ and s_{ref}. Averaging over the fixed set of random seeds stabilizes the outer objective, whereas the prescribed seed for final fitting produces one reproducible final model. The forecasting data is not used in hyperparameter optimization or coefficient estimation.

The corresponding workflow is shown in Figure 3. It separates the complete modeling sequence from the untouched forecasting sequence, makes the evaluation loop over the fixed set of random seeds explicit, and distinguishes Bayesian optimization from final model fitting.

Algorithm 3. Bayesian optimization of coupled CFSRGRVFL hyperparameters.

Input: Complete modeling sequence $\boldsymbol{T}_{\mathrm{tr}}^{(0)}$, search space Ω, fixed set of random seeds $\mathcal{S}$, prescribed seed $s_{\mathrm{ref}} \in \mathcal{S}$ for final model fitting, initial design size n_0, evaluation budget N_{BO}, and safeguards $\varepsilon_\sigma, \varepsilon_{\mathrm{MAPE}} > 0$, with $1 \le n_0 \le N_{\mathrm{BO}}$.

Output: Incumbent hyperparameter vector $\widehat{\boldsymbol{\xi}}$ and the final fitted forecasting model.

Generate an initial design $\{\boldsymbol{\xi}^{(i)}\}_{i=1}^{n_0} \subset \Omega$;
for $i = 1, \ldots, n_0$ **do**
    Compute $\sigma(\alpha, \kappa_\sigma)$ and evaluate $E_m(\boldsymbol{\xi}^{(i)})$ by solving the inner problem for every $s_m \in \mathcal{S}$;
    Set $\mathcal{J}(\boldsymbol{\xi}^{(i)}) = M^{-1} \sum_{m=1}^{M} E_m(\boldsymbol{\xi}^{(i)})$;
Set $\mathcal{D}_{n_0} = \{(\boldsymbol{\xi}^{(i)}, \mathcal{J}(\boldsymbol{\xi}^{(i)}))\}_{i=1}^{n_0}$;
for $t = n_0, \ldots, N_{\mathrm{BO}} - 1$ **do**
    Fit a probabilistic surrogate to $\mathcal{D}_t$;
    Select $\boldsymbol{\xi}^{(t+1)} \in \arg\max_{\xi\in\Omega} \mathrm{EI}_t(\boldsymbol{\xi})$;
    Compute $\sigma(\alpha, \kappa_\sigma)$, solve the inner problem for all seeds in $\mathcal{S}$, and evaluate $\mathcal{J}(\boldsymbol{\xi}^{(t+1)})$;
    Update $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(\boldsymbol{\xi}^{(t+1)}, \mathcal{J}(\boldsymbol{\xi}^{(t+1)}))\}$;
Choose $i^\star \in \arg\min_{1\le i\le N_{\mathrm{BO}}} \mathcal{J}(\boldsymbol{\xi}^{(i)})$ and set $\widehat{\boldsymbol{\xi}} = \boldsymbol{\xi}^{(i^\star)}$;
Fit the final CFSRGRVFL model on the complete modeling sequence using $\widehat{\boldsymbol{\xi}}$ and s_{ref}.

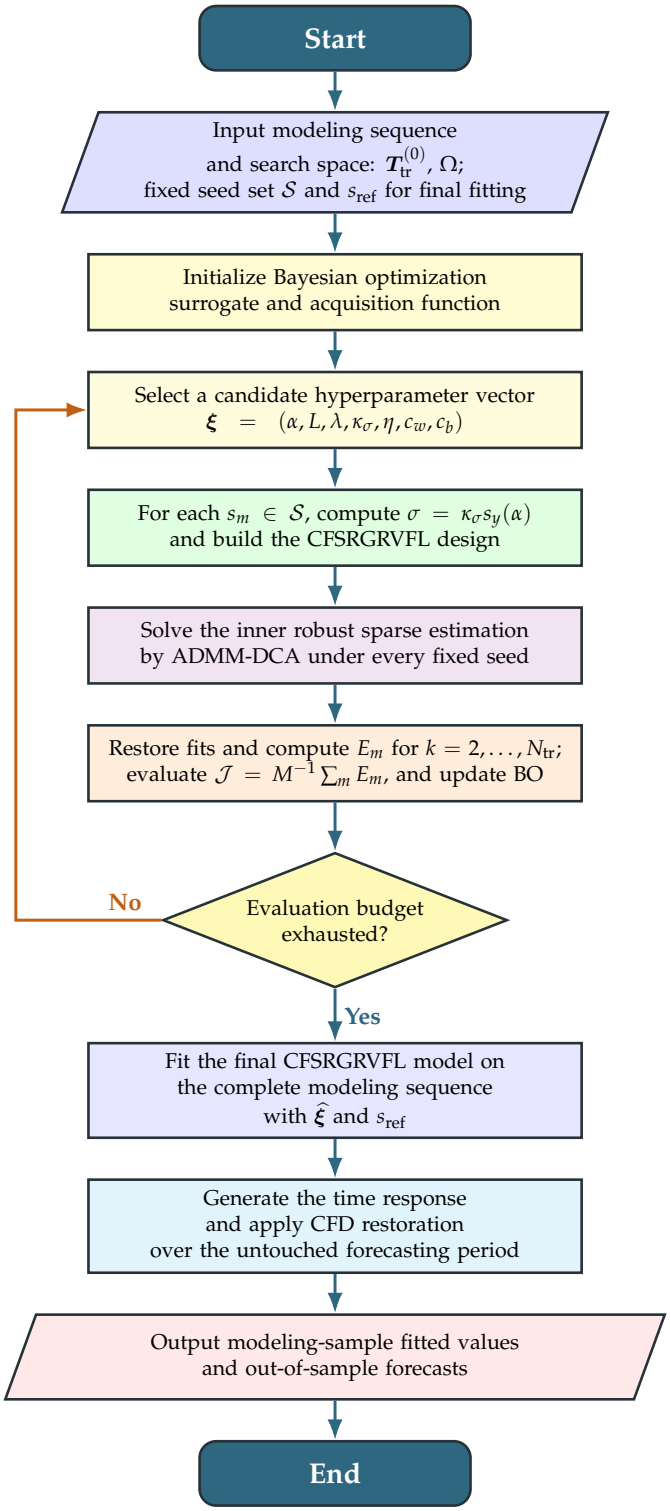


Figure 3. CFSRGRVFL workflow for Bayesian hyperparameter optimization and final model fitting.

6. Numerical experiments

This section evaluates whether the proposed CFSRGRVFL model improves both numerical accuracy and trajectory agreement for small-sample energy time series. Five empirical sequences are used to span different sample lengths, forecasting horizons, data scales, and evolution patterns. The proposed model is compared with three representative neural grey benchmarks, namely, GM-ResNet, GM-ELM, and NGSM. The evaluation is designed to answer three questions: whether CFSRGRVFL can reduce forecasting error under short modeling windows, whether it can preserve trajectory agreement when the series changes direction, and whether its advantage remains stable across heterogeneous energy indicators. The empirical analysis uses datasets derived from the *Statistical Review of World Energy 2025* [40]. For every case, the modeling sample is retained intact during hyperparameter optimization and coefficient estimation. Hyperparameters are optimized separately for each sequence because the CFA transformation and response scale are data dependent, while the same search-space construction, objective definition, and Bayesian optimization budget are used across cases. The forecasting sample is used only after the final model has been fitted on the complete modeling sample.

6.1. Evaluation metrics

To evaluate out-of-sample forecasting accuracy and agreement with the observed trajectories, nine statistical evaluation metrics are computed. The absolute-error criteria are root mean square error (RMSE), mean absolute error (MAE), and mean squared error (MSE). The relative-error criteria are normalized root mean square error (NRMSE),MAPE, root mean square percentage error (RMSPE), and relative root mean square error (RRMSE). Trajectory agreement is summarized by Willmott's index of agreement (IA), and scale-free forecast inequality is summarized by Theil's inequality coefficient (U1). The mathematical definitions are summarized in Table 3. In Table 3, $x^{(0)}(m)$ and $\widehat{x}^{(0)}(m)$ denote the observed and predicted values of the evaluated sequence. The symbol n_{ev} denotes the number of observations in the evaluated window, namely, the fitting period for the in-sample evaluation and the forecasting horizon for the out-of-sample evaluation, and $\bar{x} = (1/n_{\text{ev}}) \sum_{m=1}^{n_{\text{ev}}} x^{(0)}(m)$ is the mean of the observed values over that window. The relative metrics NRMSE, MAPE, and RMSPE are expressed in percent through the common factor 100 in their definitions.

Table 3. Definitions and mathematical formulas of the nine evaluation metrics.

Metric name	Mathematical definition
RMSE	$((1/n_{\text{ev}}) \sum_{m=1}^{n_{\text{ev}}} \left(x^{(0)}(m) - \widehat{x}^{(0)}(m)\right)^2)^{1/2}$
MAE	$(1/n_{\text{ev}}) \sum_{m=1}^{n_{\text{ev}}} \lvert x^{(0)}(m) - \widehat{x}^{(0)}(m)\rvert$
NRMSE	$(1/\bar{x})((1/n_{\text{ev}}) \sum_{m=1}^{n_{\text{ev}}} \left(x^{(0)}(m) - \widehat{x}^{(0)}(m)\right)^2)^{1/2} \times 100$
MAPE	$(1/n_{\text{ev}}) \sum_{m=1}^{n_{\text{ev}}} \lvert(x^{(0)}(m) - \widehat{x}^{(0)}(m))/x^{(0)}(m)\rvert \times 100$
RMSPE	$((1/n_{\text{ev}}) \sum_{m=1}^{n_{\text{ev}}} ((x^{(0)}(m) - \widehat{x}^{(0)}(m))/x^{(0)}(m))^2)^{1/2} \times 100$
MSE	$(1/n_{\text{ev}}) \sum_{m=1}^{n_{\text{ev}}} (x^{(0)}(m) - \widehat{x}^{(0)}(m))^2$
IA	$1 - (\sum_{m=1}^{n_{\text{ev}}} (x^{(0)}(m) - \widehat{x}^{(0)}(m))^2)/(\sum_{m=1}^{n_{\text{ev}}} (\lvert x^{(0)}(m) - \bar{x}\rvert + \lvert\widehat{x}^{(0)}(m) - \bar{x}\rvert)^2)$
U1	$(\sum_{m=1}^{n_{\text{ev}}} (x^{(0)}(m) - \widehat{x}^{(0)}(m))^2)^{1/2}/((\sum_{m=1}^{n_{\text{ev}}} x^{(0)}(m)^2)^{1/2} + (\sum_{m=1}^{n_{\text{ev}}} \widehat{x}^{(0)}(m)^2)^{1/2})$
RRMSE	$(\sum_{m=1}^{n_{\text{ev}}} (x^{(0)}(m) - \widehat{x}^{(0)}(m))^2)^{1/2}/(\sum_{m=1}^{n_{\text{ev}}} x^{(0)}(m)^2)^{1/2}$

The empirical evaluation configurations, including modeling sample sizes (N_{tr}) and forecasting horizons (n_{ev}), are summarized in Table 4.

Table 4. Data division and forecasting settings for the empirical cases.

Case	Sequence	Full sample	Modeling sample	Forecasting sample
1	Total Middle East energy supply (Mtoe)	2010–2024 ($N = 15$)	2010–2018 ($N_{\text{tr}} = 9$)	2019–2024 ($n_{\text{ev}} = 6$)
2	Total South and Central America energy supply (Mtoe)	2010–2024 ($N = 15$)	2010–2018 ($N_{\text{tr}} = 9$)	2019–2024 ($n_{\text{ev}} = 6$)
3	Saudi Arabia energy supply (Mtoe)	2010–2024 ($N = 15$)	2010–2017 ($N_{\text{tr}} = 8$)	2018–2024 ($n_{\text{ev}} = 7$)
4	Total South and Central America oil production (million tonnes)	2007–2024 ($N = 18$)	2007–2021 ($N_{\text{tr}} = 15$)	2022–2024 ($n_{\text{ev}} = 3$)
5	Total Asia Pacific CO_2 equivalent emissions (million tonnes CO_2 eq.)	2005–2024 ($N = 20$)	2005–2019 ($N_{\text{tr}} = 15$)	2020–2024 ($n_{\text{ev}} = 5$)

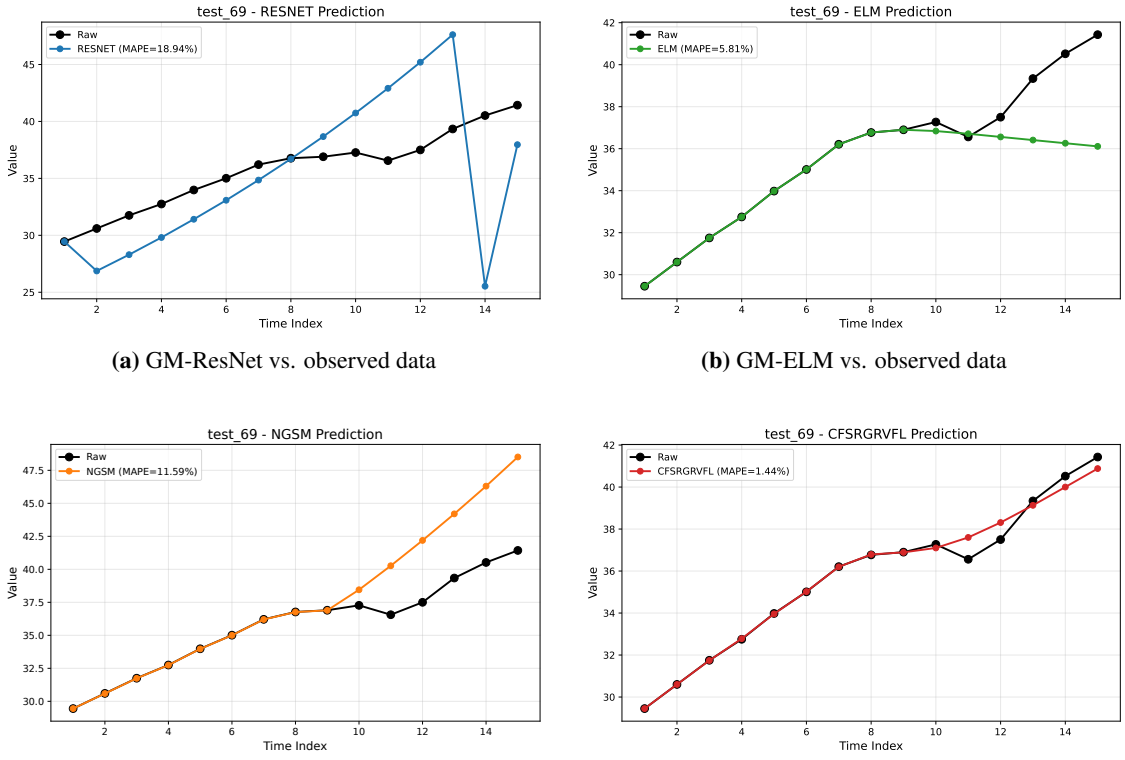
6.2. Case 1: Forecasting total energy supply I (Total Middle East)

The first evaluation uses total energy supply sequence I (Total Middle East). The modeling sample contains the first nine observations, and the subsequent six periods are used to assess out-of-sample forecasting performance. The predicted values of the compared neural grey models are reported in Table 5.

Table 5. Predicted values and trajectory fitting results for Case 1.

Year	Observed data	GM-ResNet	NGSM	GM-ELM	CFSRGRVFL
Fitting period					
2010	29.45	29.45	29.45	29.45	29.45
2011	30.60	26.87	30.59	30.60	30.60
2012	31.75	28.30	31.74	31.75	31.74
2013	32.75	29.81	32.74	32.75	32.78
2014	33.98	31.41	33.97	33.98	33.96
2015	35.01	33.08	35.00	35.01	35.02
2016	36.21	34.85	36.20	36.20	36.20
2017	36.77	36.71	36.76	36.77	36.79
2018	36.90	38.67	36.89	36.90	36.89
Forecasting period					
2019	37.27	40.74	38.45	36.84	37.10
2020	36.56	42.91	40.27	36.71	37.60
2021	37.50	45.20	42.19	36.56	38.31
2022	39.34	47.62	44.20	36.41	39.13
2023	40.52	25.53	46.30	36.26	40.00
2024	41.43	37.96	48.51	36.11	40.88

Figure 4 compares the fitted and forecast trajectories of the four models in Case 1.



(a) GM-ResNet vs. observed data

(b) GM-ELM vs. observed data

(c) NGSM vs. observed data

(d) CFSRGRVFL vs. observed data

Figure 4. Fitting and forecasting trajectories of the compared neural grey models in Case 1.

The in-sample and out-of-sample metrics for Case 1 are presented in Table 6. Over the out-of-sample period, CFSRGRVFL ranks first on all nine metrics. Its forecasting MAPE is 1.44%, which is 75.2% lower than the strongest competing result in this case (GM-ELM, 5.81%). The distinguishing feature of this case is that the modeling window ends in a flattening phase (2017-2018), after which the series resumes its growth over the six-year horizon. GM-ELM extends the flat segment into a slow decline and NGSM overshoots with nearly linear growth, whereas CFSRGRVFL follows the renewed rise and attains the highest IA (0.9761) together with the lowest U1 and RRMSE values.

Table 6. In-sample and out-of-sample evaluation metrics of the compared models in Case 1.

Model	RMSE	MAE	NRMSE (%)	MAPE (%)	RMSPE (%)	MSE	IA	U1	RRMSE
In-sample fit									
GM-ResNet	1.2580	0.9856	3.8260	2.8520	2.9030	1.5826	0.9126	0.0186	0.0378
GM-ELM	0.0520	0.0410	0.1580	0.1160	0.1190	0.0027	0.9991	0.0008	0.0016
NGSM	0.0460	0.0380	0.1390	0.1080	0.1120	0.0021	0.9994	0.0007	0.0014
CFSRGRVFL	**0.0380**	**0.0310**	**0.1150**	**0.0920**	**0.0950**	**0.0014**	**0.9997**	**0.0006**	**0.0011**
Out-of-sample forecast									
GM-ResNet	6.2899	5.9545	16.1228	18.9400	19.3715	39.5628	-0.7234	0.0771	0.1609
GM-ELM	2.4970	2.2634	6.4012	5.8100	6.3102	6.2348	0.4657	0.0323	0.0639
NGSM	4.9042	4.5500	12.6500	11.5900	12.3700	24.0510	0.0986	0.0596	0.1264
CFSRGRVFL	**0.5297**	**0.5109**	**1.3580**	**1.4400**	**1.3954**	**0.2806**	**0.9761**	**0.0068**	**0.0136**

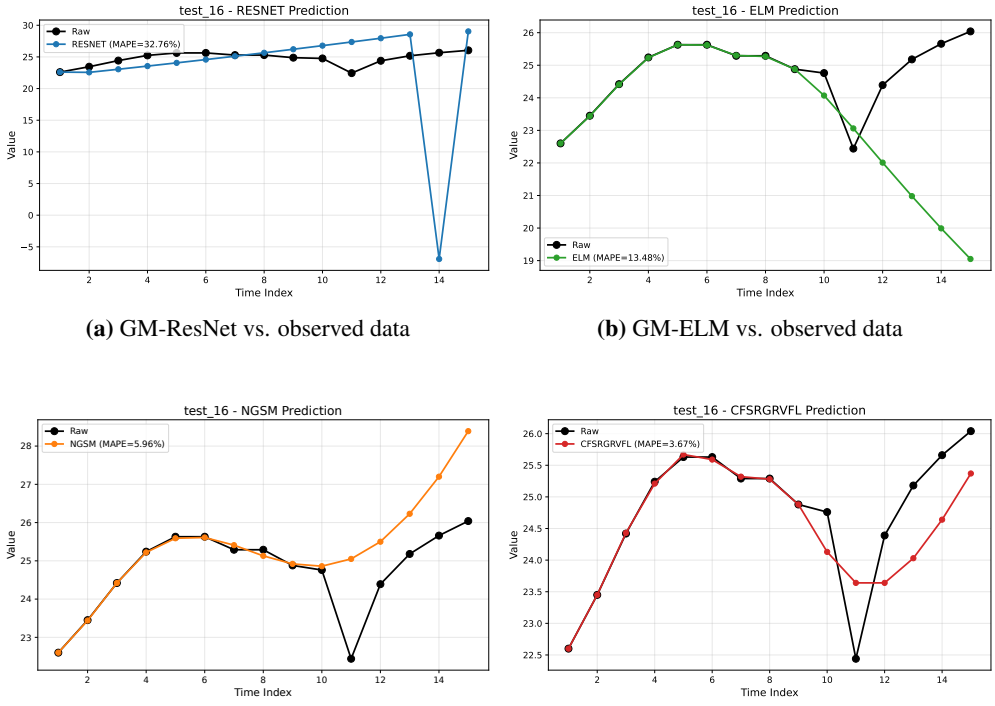
6.3. *Case 2: Forecasting total energy supply II (Total South and Central America)*

All models are evaluated on total energy supply sequence II (Total South and Central America) with $N_{\text{tr}} = 9$ in-sample observations. The observed values and model forecasts are reported in Table 7.

Table 7. Predicted values and trajectory fitting results for Case 2.

Year	Observed data	GM-ResNet	NGSM	GM-ELM	CFSRGRVFL
			Fitting period		
2010	22.60	22.60	22.60	22.60	22.60
2011	23.45	22.56	23.44	23.44	23.45
2012	24.42	23.05	24.42	24.41	24.43
2013	25.24	23.55	25.22	25.23	25.21
2014	25.63	24.06	25.59	25.63	25.67
2015	25.63	24.58	25.61	25.62	25.59
2016	25.29	25.11	25.41	25.30	25.32
2017	25.29	25.65	25.13	25.27	25.28
2018	24.88	26.21	24.92	24.88	24.88
			Forecasting period		
2019	24.76	26.78	24.86	24.07	24.13
2020	22.44	27.36	25.05	23.06	23.64
2021	24.39	27.95	25.50	22.01	23.64
2022	25.18	28.56	26.23	20.98	24.03
2023	25.66	-6.92	27.20	19.99	24.64
2024	26.04	29.03	28.39	19.05	25.37

Figure 5 compares the fitted and forecast trajectories of the four models in Case 2.



(a) GM-ResNet vs. observed data

(b) GM-ELM vs. observed data

(c) NGSM vs. observed data

(d) CFSRGRVFL vs. observed data

Figure 5. Fitting and forecasting trajectories of the compared neural grey models in Case 2.

The evaluation metrics are reported in Table 8. CFSRGRVFL provides the best out-of-sample performance on all nine metrics, with a forecasting MAPE of 3.67% against 5.96% for NGSM, a reduction of 38.4%. The distinctive pattern of this case is the dip and recovery around the sample split: the series turns downward near the end of the modeling window and rises again during the forecasting period. GM-ELM extends the decline throughout the horizon, and GM-ResNet even produces an implausible negative forecast in 2023, whereas CFSRGRVFL turns upward with the observed sequence, remains within its empirical range, and reaches an IA of 0.9262.

Table 8. In-sample and out-of-sample evaluation metrics of the compared models in Case 2.

Model	RMSE	MAE	NRMSE (%)	MAPE (%)	RMSPE (%)	MSE	IA	U1	RRMSE
In-sample fit									
GM-ResNet	0.4280	0.3540	1.7340	1.3520	1.3850	0.1832	0.9238	0.0087	0.0175
GM-ELM	0.0890	0.0710	0.3600	0.2810	0.2840	0.0079	0.9992	0.0018	0.0036
NGSM	0.0570	0.0460	0.2310	0.1790	0.1850	0.0032	0.9996	0.0012	0.0023
CFSRGRVFL	**0.0430**	**0.0350**	**0.1740**	**0.1360**	**0.1390**	**0.0018**	**0.9998**	**0.0009**	**0.0018**
Out-of-sample forecast									
GM-ResNet	13.5568	8.7362	54.7191	32.7600	57.1504	183.7869	-6.8937	0.3172	0.5486
GM-ELM	4.4226	3.5267	17.8504	13.4800	17.6448	19.5594	-0.6428	0.0880	0.1789
NGSM	1.6858	1.4600	6.8100	5.9600	6.9600	2.8420	0.4485	0.0330	0.0681
CFSRGRVFL	**0.9381**	**0.8862**	**3.7857**	**3.6700**	**3.8281**	**0.8800**	**0.9262**	**0.0188**	**0.0379**

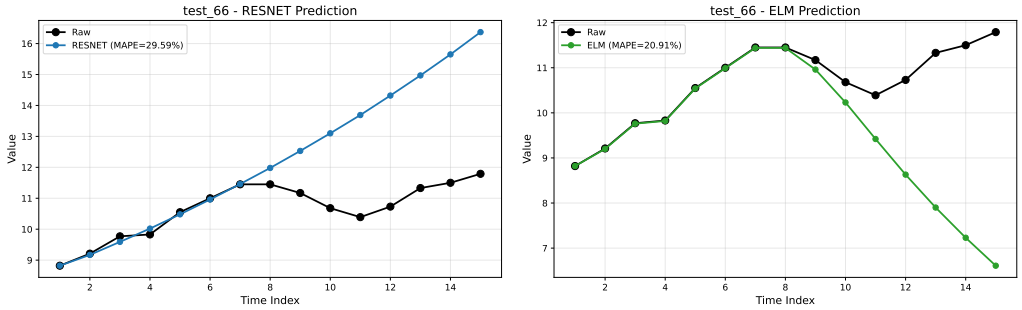
6.4. Case 3: Forecasting total energy supply III (Saudi Arabia)

Total energy supply sequence III (Saudi Arabia) is modeled using a shorter in-sample window ($N_{\text{tr}} = 8$). The observed and predicted trajectories for the out-of-sample periods are reported in Table 9.

Table 9. Predicted values and trajectory fitting results for Case 3.

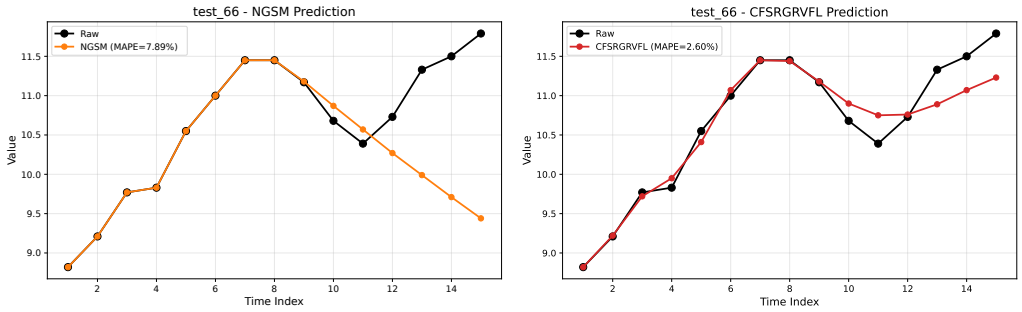
Year	Observed data	GM-ResNet	NGSM	GM-ELM	CFSRGRVFL
Fitting period					
(lr)1-6 2010	8.82	8.82	8.82	8.82	8.82
2011	9.21	9.17	9.21	9.20	9.22
2012	9.77	9.59	9.77	9.76	9.72
2013	9.83	10.02	9.83	9.82	9.95
2014	10.55	10.48	10.55	10.54	10.41
2015	11.00	10.96	11.00	10.99	11.07
2016	11.45	11.46	11.45	11.44	11.45
2017	11.45	11.98	11.45	11.44	11.44
Forecasting period					
(lr)1-6 2018	11.17	12.53	11.18	10.96	11.18
2019	10.68	13.10	10.87	10.23	10.90
2020	10.39	13.69	10.57	9.42	10.75
2021	10.73	14.32	10.27	8.63	10.76
2022	11.33	14.97	9.99	7.90	10.89
2023	11.50	15.65	9.71	7.23	11.07
2024	11.79	16.37	9.44	6.61	11.23

Figure 6 compares the fitted and forecast trajectories of the four models in Case 3.



(a) GM-ResNet vs. observed data

(b) GM-ELM vs. observed data



(c) NGSM vs. observed data

(d) CFSRGRVFL vs. observed data

Figure 6. Fitting and forecasting trajectories of the compared neural grey models in Case 3.

Table 10 reports the corresponding metrics. The defining feature of this case is the modeling window itself: Only eight observations are available for fitting, the shortest window among the five cases, and the seven-step horizon amplifies any error accumulated early in the forecast. All three competing models end with negative out-of-sample IA values, which indicate that their forecasts drift away from the observed trajectory, whereas CFSRGRVFL maintains an IA of 0.9484 and achieves the lowest forecasting MAPE (2.60%, a 67.0% reduction relative to NGSM at 7.89%).

Table 10. In-sample and out-of-sample evaluation metrics of the compared models in Case 3.

Model	RMSE	MAE	NRMSE (%)	MAPE (%)	RMSPE (%)	MSE	IA	U1	RRMSE
In-sample fit									
GM-ResNet	0.1880	0.1500	1.6370	1.4840	1.5120	0.0353	0.9856	0.0082	0.0163
GM-ELM	0.0520	0.0410	0.4530	0.4100	0.4180	0.0027	0.9991	0.0022	0.0045
NGSM	0.0480	0.0380	0.4180	0.3800	0.3890	0.0023	0.9993	0.0021	0.0042
CFSRGRVFL	**0.0380**	**0.0310**	**0.3310**	**0.2900**	**0.3010**	**0.0014**	**0.9996**	**0.0017**	**0.0033**
Out-of-sample forecast									
GM-ResNet	3.3964	3.1495	31.1139	29.5900	31.5685	11.5354	-5.7723	0.1347	0.3111
GM-ELM	2.6808	2.2942	24.5581	20.9100	24.9617	7.1867	-3.1907	0.1165	0.2456
NGSM	1.2423	0.9029	11.2100	7.8900	10.7200	1.5430	-0.5690	0.0580	0.1120
CFSRGRVFL	**0.2975**	**0.2890**	**2.7251**	**2.6000**	**2.7703**	**0.0885**	**0.9484**	**0.0136**	**0.0273**

6.5. Case 4: Forecasting oil production in Total South and Central America

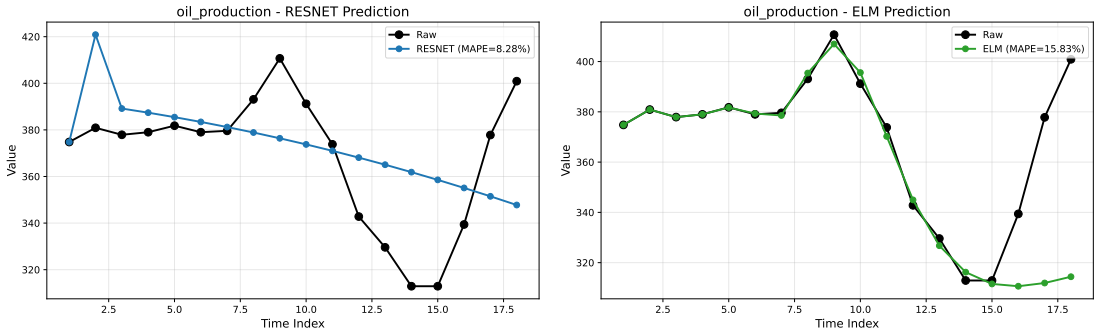
Case 4 considers the "Total South and Central America" oil-production series from the *Statistical Review of World Energy*. The dataset spans 2007 to 2024, covering crude oil, shale oil, oil sands, condensates, and natural gas liquids (NGLs). The modeling sample contains the first fifteen observations (2007–2021), and the subsequent three periods (2022–2024) are used to assess out-of-sample forecasting performance. The predicted values of the compared neural grey models are reported in Table 11.

Figure 7 compares the fitted and forecast trajectories of the four models in Case 4.

The evaluation metrics for Case 4 are reported in Table 12. This case isolates the effect of a trend reversal: the sequence declines for several years before the forecasting interval and rebounds sharply during 2022–2024, so a model that merely extends the preceding movement is unlikely to capture the rebound. The competing models either continue the earlier decline or remain near the trough level, whereas CFSRGRVFL turns upward with the observed rebound. It ranks first on all nine out-of-sample metrics and reduces the strongest competing MAPE (GM-ResNet, 8.28%) to 3.60%, a reduction of 56.5%.

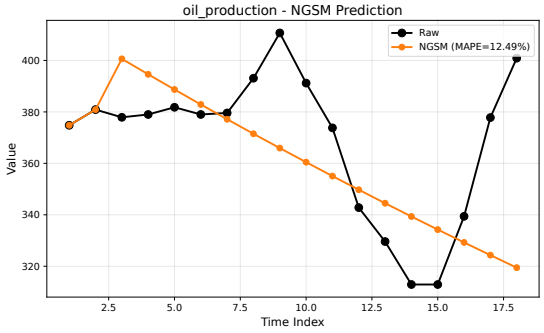
Table 11. Predicted values and trajectory fitting results for Case 4 (Total South and Central America oil production, million tonnes).

Year	Observed data	GM-ResNet	NGSM	GM-ELM	CFSRGRVFL
Fitting period					
(lr)1-6 2007	374.80	374.80	374.80	374.80	374.80
2008	380.90	420.87	380.89	380.89	380.90
2009	377.90	389.16	400.63	377.89	377.93
2010	379.00	387.39	394.62	379.00	378.99
2011	381.80	385.47	388.71	381.72	381.34
2012	379.00	383.42	382.89	379.32	380.17
2013	379.60	381.22	377.15	378.55	378.35
2014	393.10	378.88	371.50	395.42	393.93
2015	410.70	376.39	365.94	406.98	410.10
2016	391.20	373.77	360.46	395.61	391.86
2017	373.80	371.01	355.06	370.24	372.77
2018	342.80	368.10	349.74	344.97	344.53
2019	329.60	365.06	344.50	326.75	327.30
2020	312.90	361.87	339.34	316.22	314.75
2021	312.90	358.55	334.25	311.54	312.28
Forecasting period					
(lr)1-6 2022	339.40	355.09	329.25	310.62	334.12
2023	377.80	351.49	324.32	311.90	379.13
2024	400.90	347.76	319.46	314.40	436.52

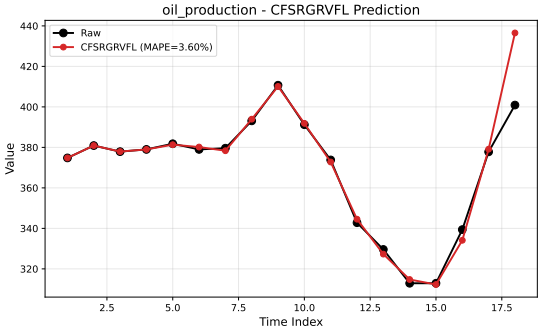


(a) GM-ResNet vs. observed data

(b) GM-ELM vs. observed data



(c) NGSM vs. observed data



(d) CFSRGRVFL vs. observed data

Figure 7. Fitting and forecasting trajectories of the compared neural grey models in Case 4.

Table 12. In-sample and out-of-sample evaluation metrics of the compared models in Case 4.

Model	RMSE	MAE	NRMSE (%)	MAPE (%)	RMSPE (%)	MSE	IA	U1	RRMSE
In-sample fit									
GM-ResNet	13.3460	9.7613	3.6120	2.7340	2.9280	178.1157	0.8746	0.0169	0.0361
GM-ELM	3.1861	2.6582	0.8620	0.7540	0.7140	10.1513	0.9924	0.0044	0.0086
NGSM	13.5969	10.5940	3.6800	2.9470	3.0090	184.8754	0.8718	0.0173	0.0368
CFSRGRVFL	**2.2685**	**1.7142**	**0.6140**	**0.4720**	**0.4960**	**5.1461**	**0.9961**	**0.0030**	**0.0061**
Out-of-sample forecast									
GM-ResNet	31.2468	26.0433	9.2920	8.2800	9.5070	976.3590	0.1819	0.0680	0.0929
GM-ELM	52.8546	45.2833	15.7180	15.8300	16.7550	2793.6040	-0.3821	0.1592	0.1572
NGSM	56.5556	48.3567	15.1800	12.4900	14.4000	3198.5000	-2.7180	0.0810	0.1514
CFSRGRVFL	**18.4599**	**14.1950**	**5.4900**	**3.6000**	**5.6490**	**340.7680**	**0.6324**	**0.0420**	**0.0549**

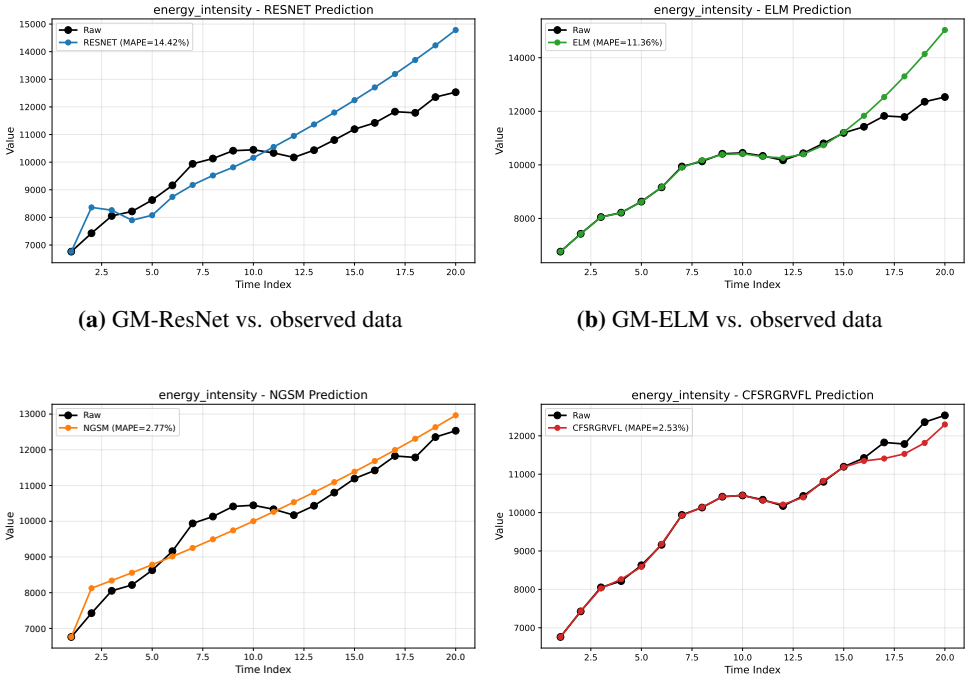
6.6. Case 5: Forecasting Total Asia Pacific CO_2-equivalent emissions

The final case considers the "Total Asia Pacific" CO_2-equivalent emissions series from the *Statistical Review of World Energy*. The dataset includes CO_2 emissions from energy combustion, flaring, methane emissions associated with fossil fuel production and distribution, and industrial process emissions. The modeling sample contains the first fifteen observations, and the subsequent five periods are used for forecasting evaluation. The trajectories are reported in Table 13.

Table 13. Predicted values and trajectory fitting results for Case 5 (Total Asia Pacific CO_2-equivalent emissions, million tonnes of CO_2 equivalent).

Year	Observed data	GM-ResNet	NGSM	GM-ELM	CFSRGRVFL
			Fitting period		
(lr)1-6 2005	6760.40	6760.40	6760.40	6760.40	6760.40
2006	7426.80	8362.15	8125.34	7423.26	7431.21
2007	8052.70	8261.79	8339.02	8048.64	8027.27
2008	8215.80	7897.47	8558.32	8212.03	8264.70
2009	8628.60	8078.14	8783.38	8620.98	8585.59
2010	9161.40	8743.04	9014.37	9168.70	9185.35
2011	9939.50	9175.20	9251.43	9901.39	9925.50
2012	10131.50	9518.12	9494.72	10171.27	10137.81
2013	10414.40	9813.53	9744.41	10390.69	10413.34
2014	10447.30	10159.17	10000.67	10416.24	10451.41
2015	10335.20	10548.74	10263.66	10305.59	10314.79
2016	10171.90	10950.18	10533.58	10260.50	10210.90
2017	10434.10	11366.13	10810.59	10398.27	10395.15
2018	10802.00	11797.17	11094.88	10730.31	10826.88
2019	11194.40	12243.90	11386.66	11221.41	11184.63
			Forecasting period		
(lr)1-6 2020	11422.00	12706.93	11686.10	11831.76	11345.48
2021	11827.20	13191.49	11993.42	12531.98	11408.29
2022	11786.50	13698.54	12308.82	13304.29	11529.44
2023	12356.30	14228.62	12632.52	14139.77	11815.32
2024	12533.40	14784.60	12964.73	15035.17	12295.46

Figure 8 compares the fitted and forecast trajectories of the four models in Case 5.



(a) GM-ResNet vs. observed data

(b) GM-ELM vs. observed data

(c) NGSM vs. observed data

(d) CFSRGRVFL vs. observed data

Figure 8. Fitting and forecasting trajectories of the compared neural grey models in Case 5.

The evaluation metrics for Case 5 are reported in Table 14. This case differs from the previous four mainly in scale, since the emissions series is substantially larger in scale than the energy-supply series, and it is also the case in which the competition is closest. CFSRGRVFL achieves the lowest MAPE (2.53%, against 2.77% for NGSM) and ranks first on eight of the nine out-of-sample metrics, while NGSM attains a slightly higher IA. For U1, the two models appear tied at the reported four-decimal precision; the unrounded values are 0.01459 for CFSRGRVFL and 0.01463 for NGSM, so CFSRGRVFL is marginally lower. Overall, CFSRGRVFL provides the most accurate forecast in this case, whereas NGSM shows slightly stronger trajectory agreement according to IA.

Table 14. In-sample and out-of-sample evaluation metrics of the compared models in Case 5.

Model	RMSE	MAE	NRMSE (%)	MAPE (%)	RMSPE (%)	MSE	IA	U1	RRMSE
In-sample fit									
GM-ResNet	659.0007	577.7860	6.9556	5.9680	6.8018	434281.9532	0.9439	0.0342	0.0689
GM-ELM	37.0540	27.4453	0.3911	0.2679	0.3577	1372.9995	0.9998	0.0019	0.0039
NGSM	421.1527	357.7007	4.4452	3.7820	4.5300	177369.5773	0.9713	0.0220	0.0440
CFSRGRVFL	**25.6954**	**20.2793**	**0.2712**	**0.2170**	**0.2812**	**660.2517**	**0.9999**	**0.0013**	**0.0027**
Out-of-sample forecast									
GM-ResNet	1774.3569	1736.9560	14.8047	14.4243	14.6629	3148342.5786	0.3757	0.0690	0.1480
GM-ELM	1575.3046	1383.5140	13.1439	11.3637	12.8138	2481584.5928	0.4520	0.0620	0.1314
NGSM	355.6794	332.0380	2.9677	2.7652	2.9597	126507.8276	**0.8568**	0.0146	0.0297
CFSRGRVFL	**345.4549**	**306.2820**	**2.8824**	**2.5339**	**2.8469**	**119339.1092**	0.8268	**0.0146**	**0.0288**

Note: the unrounded out-of-sample U1 values are 0.01459 for CFSRGRVFL and 0.01463 for NGSM; the two entries coincide only because of the four-decimal rounding used in the table.

6.7. Overall comparison and discussion

To summarize the forecasting performance across the five cases, the forecasting MAPE (%) values of all compared models are reported in Table 15. CFSRGRVFL achieves the lowest MAPE in all five cases and outperforms the competing models on most of the reported out-of-sample metrics.

Table 15. Forecasting MAPE (%) comparison across the five cases.

Case	GM-ResNet	GM-ELM	NGSM	**CFSRGRVFL**
Case 1	18.94	5.81	11.59	**1.44**
Case 2	32.76	13.48	5.96	**3.67**
Case 3	29.59	20.91	7.89	**2.60**
Case 4	8.28	15.83	12.49	**3.60**
Case 5	14.42	11.36	2.77	**2.53**

Table 16 further quantifies the cross-case advantage. The average forecasting MAPE of CFSRGRVFL is 2.77%, whereas the average MAPE values of GM-ResNet, GM-ELM, and NGSM are 20.80%, 13.48%, and 8.14%, respectively. Thus, across the five tested energy series, the proposed model reduces the average forecasting MAPE by 86.69%, 79.46%, and 65.98% compared with these three benchmarks. If the strongest non-CFSRGRVFL model is selected separately for each case, the average competing MAPE is 6.14%; CFSRGRVFL reduces this value by 54.9%. The proposed model also ranks first on most out-of-sample metrics across the five cases, confirming the consistency of its improvement across absolute-error, relative-error, trajectory-agreement, and scale-free inequality

indicators. The only exception is the IA value in Case 5, where NGSM remains slightly higher, as noted in Section 6.6.

Table 16. Average forecasting MAPE and relative reduction achieved by CFSRGRVFL.

Model	Average forecasting MAPE (%)	Reduction by CFSRGRVFL (%)
GM-ResNet	20.80	86.69
GM-ELM	13.48	79.46
NGSM	8.14	65.98
CFSRGRVFL	**2.77**	–

6.8. Ablation study

To quantify the individual contributions of the three principal components of CFSRGRVFL, a component-wise ablation study is conducted on the five forecasting cases. The full model is compared with three reduced variants: (i) w/o CFA, in which the fractional order is fixed at $\alpha = 1$ so that the order-dependent temporal weighting is removed; (ii) w/o bounded loss, in which the bounded exponential-squared loss is replaced by the ordinary squared loss; and (iii) *w/o* ℓ_1/ℓ_2, in which the norm-ratio regularization is removed. For each reduced variant, the remaining hyperparameters are re-optimized by the same Bayesian optimization protocol as the full model, and the complete selection-and-fitting pipeline is independently repeated three times; each entry of Table 17 reports the mean ± standard deviation of the forecasting MAPE over the three repetitions. The first row of Table 17 restates the full-model results of Table 15 for direct comparison.

Table 17. Ablation study: out-of-sample forecasting MAPE (%) of the full model and the reduced variants on the five cases. Variant entries report the mean ± standard deviation over three independent repetitions of the complete protocol; the full-model values are the single-run results of Table 15.

Variant	Case 1	Case 2	Case 3	Case 4	Case 5
Full CFSRGRVFL	1.44	3.67	2.60	3.60	2.53
w/o CFA (α=1)	41.02±24.01	8.57±0.65	diverged[a]	77.59±16.02	41.22±6.86
w/o bounded loss	17.37±1.15	10.97±2.61	diverged[a]	55.68±25.06	8.24±6.17
w/o ℓ_1/ℓ_2	17.26±2.75	11.08±3.49	diverged[a]	72.10±23.23	12.52±7.86

[a] Note: In Case 3, forecasting MAPE ranged from 31.5% to 3568% across the nine reduced-model fits, so the entries are marked "diverged".

The ablation results show that:

- Removing CFA increases the MAPE in all numerically stable cases (1, 2, 4, and 5) and causes divergence in Case 3, providing the most consistent evidence for the contribution of order-dependent temporal weighting.
- Replacing the bounded loss with the squared loss degrades the forecast in all stable cases (Cases 1, 2, 4, and 5), confirming its role in limiting the influence of large residuals.
- Removing the ℓ_1/ℓ_2 regularization increases the MAPE in all stable cases (Cases 1, 2, 4, and 5), demonstrating the benefit of sparse hidden-feature selection.

- In Case 3 (only eight observations), all ablated variants diverged—their multi-step forecasts did not remain bounded but grew explosively or became severely distorted (MAPE ranged from 31.5% to 3568%), so these values reflect numerical instability rather than accuracy. The dominant mechanism was a negative development coefficient causing anti-damped recursion; in the remaining fits, excessively large hidden-output coefficients produced severely distorted forecasts. In comparison, the Case 4 variants remain bounded but degrade markedly, reflecting the sensitivity of the reduced models to trend reversal.

No single component is sufficient by itself. The complete architecture provides the most consistent and stable performance across the five cases, while the contributions of the bounded loss and the ℓ_1/ℓ_2 regularization are case-dependent. The experimental discussion has been revised so that the component-specific conclusions rest on this ablation evidence rather than on the external benchmark comparison alone.

6.9. Parameter sensitivity analysis

A one-at-a-time sensitivity analysis is performed for the regularization parameter λ and the bounded-loss scale factor κ_σ (which defines $\sigma = \kappa_\sigma \,\mathrm{std}(T^{(0)})$). For each case, the hyperparameters are first selected by the same Bayesian optimization protocol; the selected values λ^* and κ_σ^* are then varied independently over eight multipliers $\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 4\}$ while all other settings are held fixed, and the forecasting MAPE is recomputed under the final fitting protocol. The MAPE is normalized by its value at the configuration selected by Bayesian optimization (BO).

Figure 9 summarizes the one-at-a-time sensitivity of forecasting MAPE to λ and κ_σ.

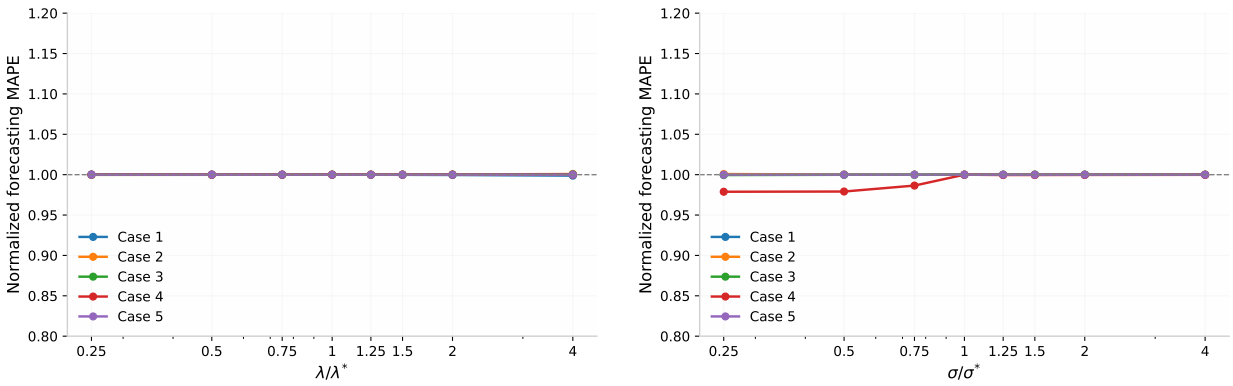


Figure 9. Sensitivity of the forecasting MAPE to (a) λ and (b) κ_σ. The horizontal axis gives the parameter ratio relative to the BO-selected value (log scale); the vertical axis gives the MAPE normalized by its value at ratio 1 (dashed line) and is drawn on a wide scale (0.8–1.2) to show that no hidden fluctuation occurs. The slight dip of Case 4 in panel (b) is associated with its trend reversal, where the loss scale modestly affects the downweighting of large residuals.

As shown in Figure 9, the normalized MAPE stays within [0.98, 1.01] for every case, multiplier, and parameter: over a sixteen-fold variation range (0.25×–4×) of either parameter, the forecasting accuracy changes by at most about 2% in relative terms, and for most settings the change is below 0.1%. The flat

response indicates that CFSRGRVFL does not rely on narrowly tuned values of λ or κ_σ; the selected values of λ^* are small (of order 10^{-3}–10^{-1}), so the estimation is dominated by the robust loss and the forecasting trajectory is insensitive to moderate perturbations of the regularization strength.

This apparent contrast with the ablation study—where removing a component is catastrophic while perturbing its strength is harmless—has a simple interpretation: the ablation removes a mechanism from the model class, whereas the sensitivity analysis varies the strength of a mechanism that is present. A small but nonzero λ still activates the ℓ_1/ℓ_2 proximal selection, so essentially the same hidden-feature support is retained across all multipliers, whereas removing the regularizer deactivates the selection altogether; likewise, every tested value of σ downweights the same large residuals, whereas the squared-loss ablation removes the downweighting entirely. The slight sensitivity of Case 4 to small σ arises from its trend reversal: the loss scale modestly affects the downweighting of the large residuals associated with the rebound, so more aggressive downweighting (smaller σ) slightly improves the forecast. For the other four cases, the BO-selected σ is large relative to the modeling residuals, so moderate multipliers do not materially change the effective shape of the loss over the range of fitted residuals, and the fitted trajectory, hence the forecast, is essentially unchanged. Together with the ablation study, these results indicate that the forecasting performance is governed by the presence of the architectural components rather than by delicate hyperparameter tuning.

7. Conclusions

This study proposed CFSRGRVFL for short energy time series forecasting. The model embeds a complete RVFL forcing term in a conformable fractional grey system: CFA and CFD control accumulation and restoration; an unpenalized direct time link preserves the RVFL skip connection; bounded exponential-squared loss limits large-residual influence; and scale-invariant ℓ_1/ℓ_2 regularization selects randomized hidden features. ADMM and Bayesian optimization handle estimation and hyperparameter tuning, respectively, without accessing the forecasting observations.

Across five energy sequences spanning different scales, window lengths, horizons, and evolution patterns, CFSRGRVFL achieved the lowest forecasting MAPE in every case and ranked first on most out-of-sample metrics. The advantage persisted in the shortest-window case and through the oil-production trend reversal. Within the tested settings, these results support the superiority of the complete formulation over GM-ResNet, GM-ELM, and NGSM.

Two additional experiments further characterize the model. A component-wise ablation showed that removing any single component markedly degrades accuracy or destabilizes multi-step extrapolation, especially under the shortest window. A sensitivity analysis over a broad parameter grid confirmed a wide stable operating region, indicating that performance is governed by the presence of the architectural components rather than by delicate parameter tuning.

Future work should include broader outlier and noise stress tests, additional independent short energy datasets, and interval or probabilistic forecasting.

Author contributions

Yang Yue conceived the study, designed the methodology, completed the program implementation, performed theoretical analysis, drafted and revised the manuscript. Shaoyong Liu collected and

processed data, carried out numerical validation and result analysis. All authors have read and approved the final manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by Talent Research Initiation Fund Project of Tongling University (2023tlxyrc56).

Conflict of interest

There are no conflicts of interest in this paper.

References

1. J. L. Deng, Control problems of grey systems, *Syst. Control Lett.*, **1** (1982), 288–294. https://doi.org/10.1016/S0167-6911(82)80025-X
2. S. Liu, Y. Lin, *Grey Systems: theory and applications*, Berlin/Heidelberg: Springer, 2010. https://doi.org/10.1007/978-3-642-16158-2
3. S. Feng, Y. Ma, Z. Song, J. Ying, Forecasting the energy consumption of China by the grey prediction model, *Energy Sources Part B Econ. Plan. Policy*, **7** (2012), 376–389. https://doi.org/10.1080/15567240903330426
4. S. B. Tsai, Using grey models for forecasting China's growth trends in renewable energy consumption, *Clean Technol. Environ. Policy*, **18** (2016), 563–571. https://doi.org/10.1007/s10098-015-1017-7
5. X. He, Y. Wang, Y. Zhang, X. Ma, W. Wu, L. Zhang, A novel structure adaptive new information priority discrete grey prediction model and its application in renewable energy generation forecasting, *Appl. Energy*, **325** (2022), 119854. https://doi.org/10.1016/j.apenergy.2022.119854
6. J. Guo, L. Wu, Y. Mu, An optimized grey model for predicting non-renewable energy consumption in China, *Heliyon*, **9** (2023), e17037. https://doi.org/10.1016/j.heliyon.2023.e17037
7. Y. Wang, X. He, Y. Zhou, Y. Luo, Y. Tang, G. Narayanan, A novel structure adaptive grey seasonal model with data reorganization and its application in solar photovoltaic power generation prediction, *Energy*, **302** (2024), 131833. https://doi.org/10.1016/j.energy.2024.131833
8. N. M. Xie, S. F. Liu, Discrete grey forecasting model and its optimization, *Appl. Math. Model.*, **33** (2009), 1173–1186. https://doi.org/10.1016/j.apm.2008.01.011
9. J. Cui, S. F. Liu, B. Zeng, N. M. Xie, A novel grey forecasting model and its optimization, *Appl. Math. Model.*, **37** (2013), 4399–4406. https://doi.org/10.1016/j.apm.2012.09.052

10. M. Cai, Non-homogeneous grey model NGM (1,1) with initial value modification and its application, *Proceedings of the 2010 2nd International Conference on Industrial and Information Systems*, 2010, 102–104. https://doi.org/10.1109/INDUSIS.2010.5565904

11. X. Ma, Z. Liu, Application of a novel time-delayed polynomial grey model to predict the natural gas consumption in China, *J. Comput. Appl. Math.*, **324** (2017), 17–24. https://doi.org/10.1016/j.cam.2017.04.020

12. L. Wu, S. Liu, L. Yao, S. Yan, D. Liu, Grey system model with the fractional order accumulation, *Commun. Nonlinear Sci. Numer. Simul.*, **18** (2013), 1775–1785. https://doi.org/10.1016/j.cnsns.2012.11.017

13. H. Duan, G. R. Lei, K. Shao, Forecasting crude oil consumption in China using a grey prediction model with an optimal fractional-order accumulating operator, *Complexity*, **2018** (2018), 3869619. https://doi.org/10.1155/2018/3869619

14. Y. Wang, P. Chi, R. Nie, X. Ma, W. Wu, B. Guo, Self-adaptive discrete grey model based on a novel fractional order reverse accumulation sequence and its application in forecasting clean energy power generation in China, *Energy*, **253** (2022), 124093. https://doi.org/10.1016/j.energy.2022.124093

15. X. Ma, W. Wu, B. Zeng, Y. Wang, X. Wu, The conformable fractional grey system model, *ISA Trans.*, **96** (2020), 255–271. https://doi.org/10.1016/j.isatra.2019.07.009

16. W. Wu, X. Ma, Y. Zhang, W. Li, Y. Wang, A novel conformable fractional non-homogeneous grey model for forecasting carbon dioxide emissions of BRICS countries, *Sci. Total Environ.*, **707** (2020), 135447. https://doi.org/10.1016/j.scitotenv.2019.135447

17. W. Wu, X. Ma, B. Zeng, P. Zhang, A conformable fractional non-homogeneous grey forecasting model with adjustable parameters CFNGMA (1,1,k,c) and its application, *J. Grey Syst.*, **36** (2024), 1–12. Available from: https://jgrey.nuaa.edu.cn/EN/Y2024/V36/I2/1.

18. W. Gong, Q. An, Novel conformable fractional order unbiased kernel regularized nonhomogeneous grey model and its applications in energy prediction, *Systems*, **13** (2025), 527. https://doi.org/10.3390/systems13070527

19. C. I. Chen, H. L. Chen, S. P. Chen, Forecasting of foreign exchange rates of Taiwan's major trading partners by novel nonlinear grey Bernoulli model NGBM (1,1), *Commun. Nonlinear Sci. Numer. Simul.*, **13** (2008), 1194–1204. https://doi.org/10.1016/j.cnsns.2006.08.008

20. J. Lu, W. Xie, H. Zhou, A. Zhang, An optimized nonlinear grey Bernoulli model and its applications, *Neurocomputing*, **177** (2016), 206–214. https://doi.org/10.1016/j.neucom.2015.11.032

21. Q. Xiao, M. Gao, X. Xiao, M. Goh, A novel grey Riccati-Bernoulli model and its application for the clean energy consumption prediction, *Eng. Appl. Artif. Intell.*, **95** (2020), 103863. https://doi.org/10.1016/j.engappai.2020.103863

22. W. Nie, H. Duan, A novel multivariable grey differential dynamic prediction model with new structures and its application to carbon emissions, *Eng. Appl. Artif. Intell.*, **122** (2023), 106174. https://doi.org/10.1016/j.engappai.2023.106174

23. X. Ma, Research on a novel kernel based grey prediction model and its applications, *Math. Probl. Eng.*, **2016** (2016), 5471748. https://doi.org/10.1155/2016/5471748
24. X. Ma, Y. Hu, Z. Liu, A novel kernel regularized nonhomogeneous grey model and its applications, *Commun. Nonlinear Sci. Numer. Simul.*, **48** (2017), 51–62. https://doi.org/10.1016/j.cnsns.2016.12.017
25. X. Ma, Z. B. Liu, The kernel-based nonlinear multivariate grey model, *Appl. Math. Model.*, **56** (2018), 217–238. https://doi.org/10.1016/j.apm.2017.12.010
26. H. Duan, D. Wang, X. Pang, Y. Liu, S. Zeng, A novel forecasting approach based on multi-kernel nonlinear multivariable grey model: a case report, *J. Clean. Prod.*, **260** (2020), 120929. https://doi.org/10.1016/j.jclepro.2020.120929
27. X. Ma, M. Xie, J. A. Suykens, A novel neural grey system model with Bayesian regularization and its applications, *Neurocomputing*, **456** (2021), 61–75. https://doi.org/10.1016/j.neucom.2021.05.048
28. Y. Hao, X. Ma, L. Song, Y. Xiang, A residual learning-based grey system model and its applications in Electricity Transformer's Seasonal oil temperature forecasting, *Eng. Appl. Artif. Intell.*, **147** (2025), 110260. https://doi.org/10.1016/j.engappai.2025.110260
29. X. Ma, Y. Hao, W. Li, An evolutionary features-based neural grey system model and its application, *Appl. Math. Model.*, **145** (2025), 116126. https://doi.org/10.1016/j.apm.2025.116126
30. Y. H. Pao, G. H. Park, D. J. Sobajic, Learning and generalization characteristics of the random vector functional-link net, *Neurocomputing*, **6** (1994), 163–180. https://doi.org/10.1016/0925-2312(94)90053-1
31. B. Igelnik, Y. H. Pao, Stochastic choice of basis functions in adaptive function approximation and the functional-link net, *IEEE Trans. Neural Netw.*, **6** (1995), 1320–1329. https://doi.org/10.1109/72.471375
32. K. K. Xu, H. X. Li, H. D. Yang, Kernel-based random vector functional-link network for fast learning of spatiotemporal dynamic processes, *IEEE Trans. Syst. Man Cybern. Syst.*, **49** (2017), 1016–1026. https://doi.org/10.1109/TSMC.2017.2694018
33. V. Kushwaha, N. M. Pindoriya, A SARIMA-RVFL hybrid model assisted by wavelet decomposition for very short-term solar PV power generation forecast, *Renew. Energy*, **140** (2019), 124–139. https://doi.org/10.1016/j.renene.2019.03.020
34. M. A. Al-qaness, A. A. Ewees, H. Fan, L. Abualigah, A. H. Elsheikh, M. Abd Elaziz, Wind power prediction using random vector functional link network with capuchin search algorithm, *Ain Shams Eng. J.*, **14** (2023), 102095. https://doi.org/10.1016/j.asej.2022.102095
35. W. Gong, G. Zong, An efficient and sparse kernelized grey RVFL network for energy forecasting, *Systems*, **14** (2026), 257. https://doi.org/10.3390/systems14030257
36. M. Wang, K. Wang, Non-convex extreme learning machine with L_1 over L_2 norm regularization and exponential squared loss for sparse and robust regression, *Signal Process.*, **244** (2026), 110524. https://doi.org/10.1016/j.sigpro.2026.110524
37. B. K. Natarajan, Sparse approximate solutions to linear systems, *SIAM J. Comput.*, **24** (1995), 227–234. https://doi.org/10.1137/S0097539792240406

38. M. Tao, Minimization of L_1 over L_2 for sparse signal recovery with convergence guarantee, *SIAM J. Sci. Comput.*, **44** (2022), A770–A797. https://doi.org/10.1137/20M136801X

39. P. I. Frazier, A tutorial on Bayesian Optimization, *arXiv*, 2018. https://arxiv.org/abs/1807.02811

40. Energy Institute, *Statistical Review of World Energy 2025*, 2025. Available from: https://www.energyinst.org/statistical-review.



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)