scientific reports

OPEN

Residual bayesian attention networks for uncertainty quantification in regression tasks

Youliang Chen[1,2], Wencan Guan[1,3,4]✉ & Rafig Azzam[3]

The demand for uncertainty quantification in modern sequence modeling tasks has prompted researchers to explore deep integration between Bayesian inference and Transformer architectures, but existing methods still face systematic engineering challenges in key technical aspects such as attention mechanism probabilization, residual connection uncertainty propagation, and epistemic-aleatoric uncertainty decoupling. This study proposes the Residual Bayesian Attention (RBA) framework, which achieves end-to-end probabilistic inference capabilities through three tightly coupled core components: Bayesian feedforward layers establish differentiable propagation mechanisms for parameter-level uncertainty, multi-layer residual Bayesian attention embeds radial basis function kernels into attention computation and introduces adaptive residual weights modeled by Beta distributions, and the Bayesian covariance construction module generates mathematically rigorous covariance representations through outer product operations and eigenvalue correction. Systematic evaluation on benchmark datasets covering six domains including engineering optimization, time series forecasting, and spatial modeling demonstrates that RBA achieves stable uncertainty quantification performance in medium-scale structured data scenarios, particularly exhibiting technical advantages in prediction interval calibration quality. Notably, through objective evaluation of challenging tasks such as complex physical systems, this study identifies the common technical boundaries of current deep learning methods in multi-physics coupled system modeling, providing important empirical insights for the development direction of this field. Therefore, RBA, as a systematic engineering integration framework of Bayesian inference and Transformer architecture, provides a methodological contribution with clearly defined applicability boundaries for principled uncertainty quantification in deep sequence modeling.

Keywords Residual bayesian attention, Uncertainty quantification, Transformer architecture, Gaussian process, Variational inference

The field of uncertainty quantification and deep learning has established a solid theoretical foundation through hybrid approaches such as scalable variational Gaussian processes, neural network-Gaussian process correspondences, and deep Gaussian processes[1–6]. While these hybrid methods have achieved significant success in their respective application domains, the widespread reliance on attention mechanisms in modern sequence modeling and multimodal tasks makes principled uncertainty quantification in Transformer architectures an important engineering requirement[7,8]. However, the systematic engineering integration of these mature Bayesian techniques with modern Transformer architectures still faces concrete technical challenges, particularly in designing effective uncertainty propagation pathways within attention mechanisms, implementing probabilistic distribution composition under residual connections, and addressing the decoupling and fusion of epistemic and aleatoric uncertainty while maintaining mathematical rigor[9–12].

In the probabilistic modeling of attention mechanisms, works such as variational attention mechanisms[13] and stochastic attention weights[14] have preliminarily explored uncertainty representations of attention scores, laying the foundation for combining attention mechanisms with probabilistic reasoning. Addressing uncertainty propagation in deep networks, Monte Carlo Dropout[15] achieves approximate sampling of parameter uncertainty

[1]Department of Civil Engineering, University of Shanghai for Science and Technology, Shanghai 200093, 516 Jungong Rd, PR China. [2]Department of Engineering Geology, School of Civil and Environmental Engineering, University of New South Wales, Kensington, NSW 2052, Australia. [3]Department of Engineering Geology and Hydrogeology, RWTH Aachen University, D-52064 Aachen, Lochnerstr. 4–20 Haus A, Germany. [4]European Scoiety for Fuzzy Logic and Technology, Mieres 33600, Manual Llaneza 32, Spain. ✉email: guanwencan123@outlook.com; wencan.guan@unsw.edu.au

by maintaining dropout activation during inference, while Bayesian Deep Learning methods[16,17] establish probabilistic distribution representations at the weight level through variational inference frameworks. In the domain of uncertainty decoupling and fusion, Deep Ensemble[18] effectively separates epistemic and aleatoric uncertainty through model ensembling, Sensoy et al. (2018)[19] provides a unified framework for uncertainty quantification based on evidence theory, and multi-task uncertainty modeling[20] makes important contributions in covariance construction. Although these works have achieved significant progress in their respective technical directions, the systematic engineering integration of these technologies in Transformer architectures, particularly in maintaining the balance between computational feasibility and theoretical consistency, still lacks a unified solution framework.

In the theoretical construction of probabilistic attention mechanisms, variational information bottleneck theory[21] provides a principled foundation for information-theoretic interpretation of attention weights, while kernelized attention mechanisms[22] achieve linear complexity approximation through the FAVOR + algorithm. Gaussian Process Attention[23] directly embeds Gaussian process priors into attention computation. In the technical evolution of fusing variational inference with deep architectures, Structured Variational Inference[24] achieves more precise posterior approximation through structured priors, Multiplicative Normalizing Flows[25] provides more expressive variational families for complex posterior distribution modeling, while Concrete Dropout[15] and Variational Dropout[26] establish bridges between theory and practice in adaptive regularization. The fusion of covariance modeling with multi-task learning reflects more sophisticated mathematical treatment: Multi-Task Gaussian Processes[27] establish rigorous mathematical representations of inter-task correlations, Tensor-Structured Gaussian Processes[28] achieve efficient computation of high-dimensional covariances through Kronecker decomposition, and Neural Module Networks[29] and Modular Meta-Learning[30] provide composable uncertainty modeling approaches in modular architecture design. Although these deep technical developments have reached considerable theoretical depth in their respective subfields, there currently lacks a unified engineering framework to systematically integrate these advanced technologies, particularly in implementing end-to-end Bayesian inference capabilities in the dominant Transformer architecture.

To address this systematic engineering integration challenge, this research proposes the Residual Bayesian Attention (RBA) framework, which achieves organic fusion of Bayesian inference with Transformer architectures through three tightly coupled core components. Specifically, the Bayesian feedforward layer establishes a differentiable propagation mechanism for parameter-level uncertainty through reparameterization tricks and Delta method approximation. The multi-layer residual Bayesian attention directly embeds radial basis function kernels into attention score computation and introduces Beta distribution-modeled adaptive residual weights to enable uncertainty accumulation propagation in deep networks. The Bayesian covariance construction module generates covariance matrix representations that satisfy Gaussian process mathematical requirements through outer product operations and eigenvalue correction techniques. This design operates synergistically under a unified variational Bayesian optimization framework, maintaining the parallel computation advantages of Transformers while achieving principled separation of epistemic and aleatoric uncertainty, providing end-to-end Bayesian inference capabilities for modern sequence modeling tasks.

In comprehensive validation across six benchmark datasets covering different application domains and data characteristics, RBA demonstrates stable and competitive performance: achieving a coefficient of determination of 0.972 and good calibration quality (ECE = 0.1877) in engineering optimization tasks, maintaining prediction accuracy of 0.920 while controlling the Prediction Interval Normalized Average Width to 0.180 in time series forecasting, and reaching 96.38% Prediction Interval Coverage Probability in spatial modeling tasks while effectively handling geospatial dependencies. Notably, RBA exhibits consistent advantages in uncertainty calibration, maintaining good consistency between prediction confidence and actual accuracy across various data complexity levels—a characteristic of practical value for decision support in real applications. Although complex physical system modeling still presents challenges[2], this objective delineation of applicability boundaries further validates the objectivity and credibility of the method evaluation. Therefore, RBA, as a systematic engineering integration framework for Bayesian inference and Transformer architectures, contributes an engineering solution with clearly defined technical boundaries for probabilistic reasoning implementation in deep sequence modeling.

Methods

RBA construction

The Residual Bayesian Attention (RBA) algorithm represents an innovative model that deeply integrates the Bayesian inference of Gaussian processes with the residual architecture of Transformers. The entire algorithmic construction adheres to the design principles of "Bayesian consistency + residual information propagation + multi-layer uncertainty accumulation."

As illustrated in Fig. 1, the core framework comprises three components: Bayesian feedforward layers, multi-layer residual-connected Bayesian attention, and Bayesian covariance construction. Specifically, the Residual Bayesian Attention (RBA) architecture achieves essential fusion between deep learning and Gaussian processes through three tightly coupled core components. The Bayesian feedforward layer constitutes the probabilistic foundation of the entire architecture, transforming deterministic parameters in traditional neural networks into probability distributions characterized by means and variances, implementing differentiable stochastic sampling through reparameterization tricks, and quantifying epistemic uncertainty at the parameter level using variational inference frameworks, thereby providing a probabilistic feature representation foundation for subsequent components. The multi-layer residual Bayesian attention mechanism builds upon this foundation to construct a deep probabilistic inference framework, directly embedding Gaussian process kernel function concepts into attention score computations, enabling each attention head to correspond to an independent Gaussian process prior, learning data correlations across different scales and patterns through multi-head

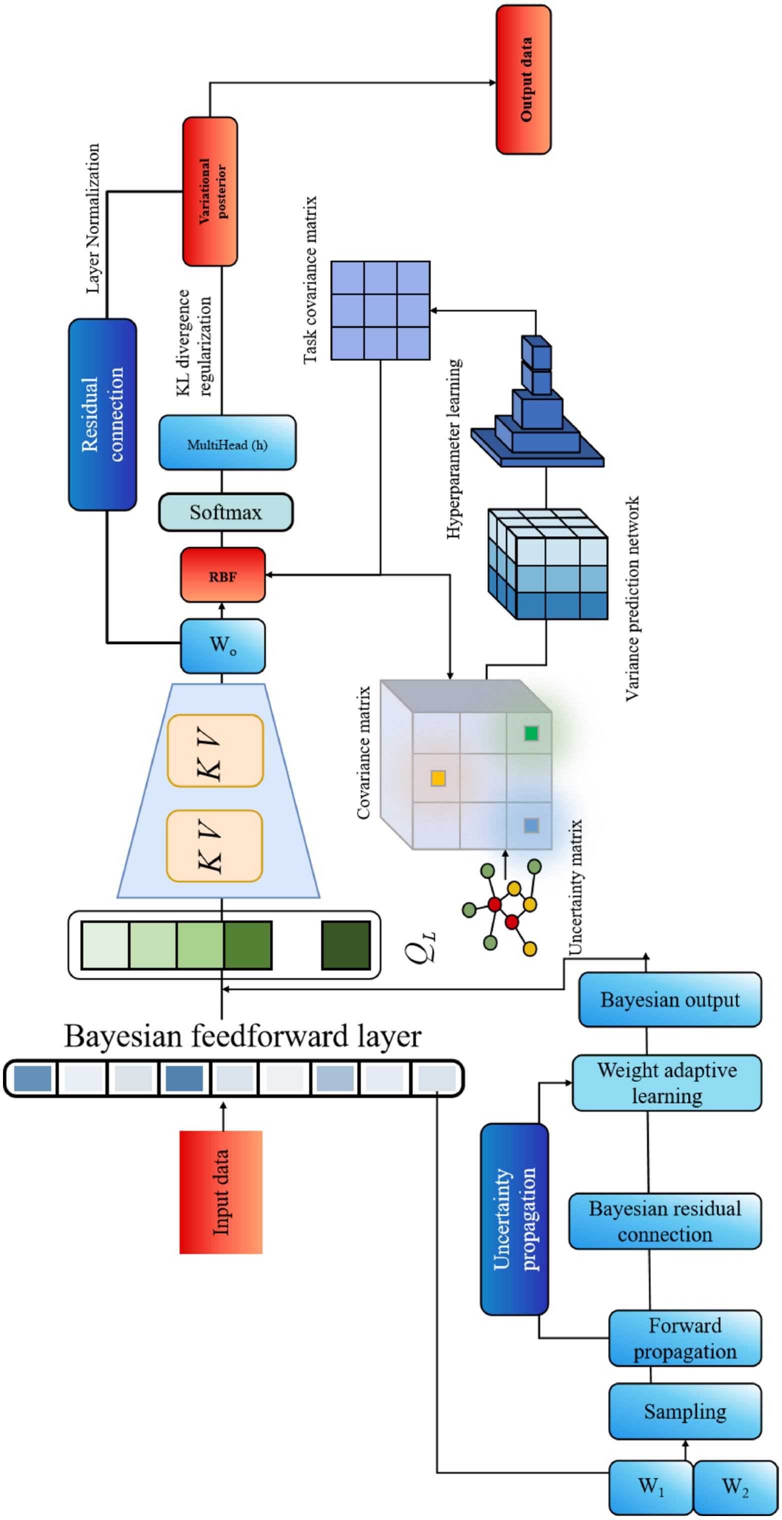


Fig. 1. Residual Bayesian Attention (RBA) architecture

parallel processing, while introducing Bayesian residual connection mechanisms to address gradient vanishing problems in deep networks and achieving effective propagation and accumulation of uncertainty in the network depth direction through hierarchical variational inference. The Bayesian covariance construction component is responsible for transforming the outputs of the aforementioned multi-layer attention mechanisms into covariance matrices that satisfy the mathematical requirements of Gaussian processes, generating fundamental covariance structures through outer product operations, separately modeling epistemic components arising from parameter uncertainty and aleatoric components from intrinsic data randomness, ensuring matrix positive definiteness through eigenvalue correction, and supporting tensor product extensions for multi-task scenarios. The three components undergo collaborative optimization within a unified variational Bayesian framework, forming an end-to-end learning system that possesses both the powerful representational capabilities of deep neural networks and maintains the rigorous uncertainty quantification characteristics of Gaussian processes, achieving organic unification of data-driven feature learning and probabilistic inference.

Bayesian feedforward layer

The Bayesian feedforward layer transforms deterministic neural network parameters into variational posterior distributions (Eq. 7) through a hierarchical Bayesian model (Eq. 6), achieving explicit modeling and differentiable propagation of parameter-level epistemic uncertainty. Equations (8)-(9) implement differentiable parameter sampling, decoupling randomness from variational parameters through the reparameterization trick to ensure numerical stability of backpropagation. Equations (13)-(26) establish a complete uncertainty propagation pathway, from expectation-variance computation of linear transformations (Eqs. 15–17) to Delta method approximation of activation functions (Eqs. 11–14), and further to distribution composition of residual connections (Eq. 24), achieving analytical propagation of inter-layer uncertainty. Equations (27)-(33) construct the Evidence Lower Bound (ELBO) framework, balancing reconstruction error with prior constraints through KL divergence regularization, and establishing Jacobian matrix representation for uncertainty propagation (Eq. 33), laying the theoretical foundation for subsequent probabilistic modeling of attention mechanisms.

The Bayesian feedforward layer process is as follows.

The first layer weights can be expressed as formulas (1) and (2).

$$W_1 \sim N(\mu_{W_1}, \Sigma_{W_1}) \tag{1}$$

$$b_1 \sim N(\mu_{b_1}, \Sigma_{b_1}) \tag{2}$$

The second layer of weight priors can be expressed as formulas (3) and (4).

$$b_2 \sim N(\mu_{b_2}, \Sigma_{b_2}) \tag{3}$$

$$W_2 \sim N(\mu_{W_2}, \Sigma_{W_2}) \tag{4}$$

Among them, the covariance matrix is a diagonal matrix defined by formula (5).

$$\Sigma_{W_i} = \text{diag}(\sigma^2_{W_i}), \quad \Sigma_{b_i} = \text{diag}(\sigma^2_{b_i}) \tag{5}$$

The hierarchical Bayesian model[31] is defined by formula (6).

$$p(\theta) = p(W_1)p(b_1)p(W_2)p(b_2) \tag{6}$$

Where $\theta = \{W_1, b_1, W_2, b_2\}$ is for all parameters.

For difficult-to-handle true posteriors, use variational posteriors defined by formula (7).

$$q(\theta) = q(W_1)q(b_1)q(W_2)q(b_2) \tag{7}$$

Each factor is $q(W_i) = N(\mu_{q,W_i}, diag(\sigma^2_{q,W_i}))$, $q(b_i) = N(\mu_{q,b_i}, diag(\sigma^2_{q,b_i}))$,which is represented by Eqs. (8) and (9) from the variational posterior[32] sampling process.

$$W_i^{(s)} = \mu_{q,W_i} + \sigma_{q,W_i} \odot O_{W_i}^{(s)} \tag{8}$$

$$b_i^{(s)} = \mu_{q,b_i} + \sigma_{q,b_i} \odot \epsilon_{b_i}^{(s)} \tag{9}$$

Among them,$\epsilon_{W_i}^{(s)}, \epsilon_{b_i}^{(s)} \sim N(0, I)$,GELU[33] function is defined as formula (10).

$$\text{GELU}(z) = z\Phi(z) = z\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{10}$$

The approximate form is $\text{GELU}(z) \approx \frac{z}{2}\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(z + 0.044715z^3\right)\right)\right)$.

For random variables $Z \sim N(\mu_Z, \sigma^2_Z)$ and differentiable functions $g(\cdot)$, the Delta method provides the solution to formula (11).

$$g(Z) \approx N\left(g(\mu_Z), \left(\frac{dg}{dz}|_{z=\mu_Z}\right)^2 \sigma^2_Z\right) \tag{11}$$

The derivative of GELU is defined by formula (12).

$$\frac{d\text{GELU}}{dz} = \Phi(z) + z\phi(z) \tag{12}$$

Among them, $\Phi(z) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{z}{\sqrt{2}}\right)\right)$ is the standard normal CDF[34],$\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ is the standard normal PDF.

Input $H_1 \sim N(\mu_{H_1}, \Sigma_{H_1})$, activation function output formula (13)-(14) value.

$$E[\text{GELU}(H_1)] = \text{GELU}(\mu_{H_1}) \tag{13}$$

$$\text{Var}[\text{GELU}(H_1)] = \left(\frac{d\text{GELU}}{dh}|_{h=\mu_{H_1}}\right)^2 \odot \text{diag}(\Sigma_{H_1}) \tag{14}$$

Bayesian forward propagation in feedforward networks, where the first-layer linear transformation can be defined by formula (15).

$$H_1^{(s)} = XW_1^{(s)} + b_1^{(s)} \tag{15}$$

The expected value of the first-layer output distribution is expressed by formula (16).

$$E[H_1] = XE[W_1] + E[b_1] = X\mu_{q,W_1} + \mu_{q,b_1} \tag{16}$$

Variance (diagonal approximation)[35] is expressed by formula (17).

$$\text{Var}[H_{1,ij}] = \sum_{k=1}^{d} X_{ik}^2 \sigma_{q,W_1,kj}^2 + \sigma_{q,b_1,j}^2 \tag{17}$$

The distribution after activation can be expressed by formula (18).

$$H_1^{\text{act}} = \text{GELU}(H_1) \tag{18}$$

Applying the Delta method[36] yields the activated distributions of Eqs. (19) and (20).

$$E[H_1^{\text{act}}] = \text{GELU}(E[H_1]) \tag{19}$$

$$\text{Var}[H_1^{\text{act}}] = \left(\frac{d\text{GELU}}{dh}|_{h=E[H_1]}\right)^2 \odot \text{Var}[H_1] \tag{20}$$

The second-order linear transformation is expressed by formula (21).

$$Y^{(s)} = H_1^{\text{act}(s)} W_2^{(s)} + b_2^{(s)} \tag{21}$$

Expectation is defined as formula (22).

$$E[Y] = E[H_1^{\text{act}}]E[W_2] + E[b_2] \tag{22}$$

The variance is given by formula (23).

$$\text{Var}[Y_{ij}] = \sum_{k=1}^{d_{ff}} E[H_1^{\text{act}}]_{ik}^2 \sigma_{q,W_2,kj}^2 + \text{Var}[H_1^{\text{act}}]_{ik}(\mu_{q,W_2,kj})^2 + \sigma_{q,b_2,j}^2 \tag{23}$$

Construct a Bayesian model with residual connections, where the probability of residual connections is represented as follows. Let the input of the layer be $X_l \sim N(\mu_{X_l}, \Sigma_{X_l})$, the output of the feedforward network be $F_l \sim N(\mu_{F_l}, \Sigma_{F_l})$, and the residual connection be $X_{l+1} = X_l + F_l$.

Since the linear combination of normal distributions is still a normal distribution, we obtain formula (24).

$$X_{l+1} \sim N(\mu_{X_l} + \mu_{F_l}, \Sigma_{X_l} + \Sigma_{F_l}) \tag{24}$$

For layer normalization transformation, update to formula (25).

$$Y = \gamma \odot \frac{X - \mu_X}{\sigma_X} + \beta \tag{25}$$

Among them, $\mu_X = \frac{1}{d}\sum_{i=1}^{d} X_i, \quad \sigma_X^2 = \frac{1}{d}\sum_{i=1}^{d}(X_i - \mu_X)^2$.

the approximate transformation of uncertainty is defined by formula (26)

$$\text{Var}[Y_i] \approx \left(\frac{\gamma_i}{\sigma_X}\right)^2 \text{Var}[X_i] \tag{26}$$

The evidence lower bound (ELBO)[37] of the variational objective function is defined by formula (27).

$$\mathcal{L} = E_{q(\theta)}[\log p(Y|X,\theta)] - KL[q(\theta)||p(\theta)] \tag{27}$$

The error term of the reconstruction is defined as formula (28).

$$E_{q(\theta)}[\log p(Y|X,\theta)] = E_{q(\theta)}\left[-\frac{1}{2\sigma^2}||Y - f(X;\theta)||^2\right] \tag{28}$$

Introducing the KL divergence term[38], each parameter is defined by formula (29).

$$KL[q(W_i)||p(W_i)] = \frac{1}{2}\left[\mathrm{tr}(\Sigma_{p,W_i}^{-1}\Sigma_{q,W_i}) + (\mu_{p,W_i} - \mu_{q,W_i})^T\Sigma_{p,W_i}^{-1}(\mu_{p,W_i} - \mu_{q,W_i}) - k + \log\frac{|\Sigma_{p,W_i}|}{|\Sigma_{q,W_i}|}\right] \tag{29}$$

Using the reparameterization technique[39], the gradient update is given by Eqs. (30) and (31).

$$\nabla_{\mu_q}\mathcal{L} = E_\epsilon[\nabla_{\mu_q}f(X;\mu_q + \sigma_q \odot \epsilon)] \tag{30}$$

$$\nabla_{\sigma_q}\mathcal{L} = E_\epsilon[\epsilon \odot \nabla_{\mu_q}f(X;\mu_q + \sigma_q \odot \epsilon)] \tag{31}$$

Given the input distribution, the output of the Bayesian feedforward layer is given by Eq. (32).

$$(Y, \Sigma_Y) = \mathrm{BayesianFF}(X, \Sigma_X; q(\theta)) \tag{32}$$

End-to-end uncertainty propagation can be determined by formula (33).

$$\Sigma_Y = J_f(\mu_X)\Sigma_X J_f(\mu_X)^T + E[\nabla_\theta f(\mu_X;\theta)\Sigma_\theta\nabla_\theta f(\mu_X;\theta)^T] \tag{33}$$

Among them, $J_f(\mu_X)$ is the Jacobian matrix[40] of the function with respect to the input, and Σ_θ is the covariance matrix of the parameters.

Multi-layer residual bayesian attention

Multi-layer Residual Bayesian Attention transforms traditional similarity measures into probabilistic correlation modeling by directly embedding Gaussian process kernel functions into attention score computation (Eqs. 41–42), where each attention head corresponds to independent Gaussian process priors (Eqs. 34–40) to learn multi-scale data correlation structures. In terms of uncertainty propagation in deep networks, while traditional residual connections address the vanishing gradient problem, the Bayesian residual connection mechanism achieves cumulative variance propagation (Eq. 52) and hierarchical variational inference (Eq. 51) through Beta distribution-modeled adaptive weights (Eqs. 46–46) and Bayesian extension of Layer Normalization (Eqs. 48–50), combined with the parameter-level uncertainty foundation established in Sect. 2.2.1, ensuring cumulative propagation of uncertainty in the depth direction of the network. The entire framework integrates parameter uncertainty from feedforward layers with correlation uncertainty from attention mechanisms through unified variational lower bound optimization (Eqs. 53–56), forming a complete deep probabilistic inference system.

The complete mathematical derivation process of the multi-layer residual Bayesian attention[41] is as follows. The prior definitions of the query, key, and value projection weights in the *l* th layer are given by formulas (34) to (37).

$$W_Q^{(l)} \sim N(\mu_{W_Q^{(l)}}, \Sigma_{W_Q^{(l)}}) \tag{34}$$

$$W_K^{(l)} \sim N(\mu_{W_K^{(l)}}, \Sigma_{W_K^{(l)}}) \tag{35}$$

$$W_V^{(l)} \sim N(\mu_{W_V^{(l)}}, \Sigma_{W_V^{(l)}}) \tag{36}$$

$$W_O^{(l)} \sim N(\mu_{W_O^{(l)}}, \Sigma_{W_O^{(l)}}) \tag{37}$$

The kernel function parameters of the nth attention head are defined by formulas (38)-(40).

$$\ell_h^{(l)} \sim \mathrm{Gamma}(\alpha_\ell, \beta_\ell) \tag{38}$$

$$\sigma_f^{(l),h} \sim \mathrm{Gamma}(\alpha_f, \beta_f) \tag{39}$$

$$\sigma_n^{(l),h} \sim \mathrm{Gamma}(\alpha_n, \beta_n) \tag{40}$$

Among them, $\ell_h^{(l)}$ is the length scale, $\sigma_f^{(l),h}$ is the signal variance, and $\sigma_n^{(l),h}$ is the noise variance.

The RBF kernel function is defined by formula (41).

$$k(x_i, x_j) = (\sigma_f^{(l),h})^2 \exp\left(-\frac{\| x_i - x_j \|^2}{2(\ell_h^{(l)})^2}\right) \tag{41}$$

The attention score of the l th layer and the h th head is expressed by formula (42).

$$S_{ij}^{(l),h} = \frac{Q_i^{(l),h} \cdot K_j^{(l),h}}{\sqrt{d_k}} + k(x_i, x_j) \tag{42}$$

Softmax normalization is expressed as formula (43).

$$A_{ij}^{(l),h} = \frac{\exp(S_{ij}^{(l),h})}{\sum_{k=1}^{n} \exp(S_{ik}^{(l),h})} \tag{43}$$

The first output is formula (44).

$$\text{head}^{(l),h} = \sum_{j=1}^{n} A_{ij}^{(l),h} V_j^{(l),h} \tag{44}$$

Multiple series connections can be represented by formula (45).

$$\text{MultiHead}^{(l)} = \text{Concat}(\text{head}^{(l),1}, \ldots, \text{head}^{(l),H}) W_O^{(l)} \tag{45}$$

The residual connection weight prior is defined by Eqs. (46) and (46).

$$\beta_{\text{att}}^{(l)} \sim \text{Beta}(\gamma_3, \gamma_4) \tag{46}$$

$$\alpha_{\text{res}}^{(l)} \sim \text{Beta}(\gamma_1, \gamma_2) \tag{47}$$

Residual update equation, extended to Bayesian formula (48)-(49) based on Layer Normalization.

$$\mu_{\text{norm}} = \frac{1}{d_h} \sum_{i=1}^{d_h} x_i \tag{48}$$

$$\sigma_{\text{norm}}^2 = \frac{1}{d_h} \sum_{i=1}^{d_h} (x_i - \mu_{\text{norm}})^2 + \epsilon \tag{49}$$

The residual connection is determined by formula (50).

$$H^{(l+1)} = \alpha_{\text{res}}^{(l)} \cdot \text{LayerNorm}(H^{(l)}) + \beta_{\text{att}}^{(l)} \cdot \text{MultiHead}^{(l)} \tag{50}$$

The variational posterior of the layer is given by formula (51).

$$q(\theta^{(l)}) = \prod_{p \in \{\text{Q,K,V,O}\}} q(W_p^{(l)}) \prod_{h=1}^{H} q(\ell_h^{(l)}) q(\sigma_f^{(l),h}) q(\sigma_n^{(l),h}) \tag{51}$$

Cumulative variance propagation is given by Eq. (52).

$$\text{Var}(H^{(l+1)}) = (\alpha_{\text{res}}^{(l)})^2 \text{Var}(H^{(l)}) + (\beta_{\text{att}}^{(l)})^2 \text{Var}(\text{MultiHead}^{(l)}) \tag{52}$$

Introducing KL divergence regularization, the variational lower bound objective function becomes Eq. (53).

$$\mathcal{L} = \mathbb{E}_{q(\theta)}[\log p(y|X, \theta)] - \sum_{l=1}^{L} KL[q(\theta^{(l)}) \parallel p(\theta^{(l)})] \tag{53}$$

$$\text{KL}[q(\theta^{(l)}) p(\theta^{(l)})] = \sum_{p \in \{\text{Q,K,V,O}\}} \text{KL}[q(W_p^{(l)}) r p(W_p^{(l)})] + \sum_{h=1}^{H} \sum_{\phi \in \{\ell, \sigma_f, \sigma_n\}} \text{KL}[q(\phi_h^{(l)}) p(\phi_h^{(l)})]$$

The complete model posterior representation is given by Eq. (54).

$$p(\Theta|X, y) \propto p(y|X, \Theta) \prod_{l=1}^{L} p(\theta^{(l)}) \tag{54}$$

Among them, $\Theta = \{\theta^{(1)}, \ldots, \theta^{(L)}\}$ is for all layer parameters.

The variational approximation is given by formula (55).

$$q(\Theta) = \prod_{l=1}^{L} q(\theta^{(l)}) \tag{55}$$

The optimal variational parameter is obtained by maximizing the variational lower bound in Eq. (56).

$$\phi^* = \arg\max_{\phi} \mathcal{L}(\phi) \tag{56}$$

Bayesian covariance construction

Bayesian covariance construction transforms the deep probabilistic inference outputs from Sect. 2.2.2 into covariance matrices that satisfy Gaussian process mathematical requirements, constructing basic covariance structures through outer product operations (Eqs. 57–58) to achieve mapping from network representations to probabilistic kernel functions. The covariance enhancement mechanism separately models epistemic uncertainty and aleatoric uncertainty: epistemic uncertainty through Jacobian matrix computation of parameter propagation (Eq. 60), and aleatoric uncertainty through independent variance prediction network modeling (Eqs. 61–63), with their fusion forming a complete uncertainty matrix (Eqs. 59, 65). The mathematical correction process ensures the mathematical validity of covariance matrices through symmetrization processing (Eq. 66), eigenvalue decomposition (Eq. 67), and positive definiteness correction (Eqs. 68–69). Multi-task extensions adopt tensor product decomposition structures (Eqs. 70–73), decoupling data correlations from task correlations in modeling. Covariance hyperparameter learning achieves adaptive adjustment of global scaling factors through variational posterior updates (Eqs. 74–78), while covariance quality assessment ensures numerical stability through condition number monitoring (Eq. 79), rank deficiency detection (Eq. 80), and Frobenius norm conditions (Eq. 81), ultimately establishing a complete mapping from network outputs to Gaussian process prediction covariance (Eqs. 83–85).

The Bayesian covariance[42] construction process is as follows.

The final layer output is reduced to formula (57).

$$Z^{(L)} = H^{(L)} W_{\text{proj}} + b_{\text{proj}} \tag{57}$$

Among them,

$$W_{\text{proj}} \sim N(\mu_{W_{\text{proj}}}, \Sigma_{W_{\text{proj}}})$$
$$b_{\text{proj}} \sim N(\mu_{b_{\text{proj}}}, \Sigma_{b_{\text{proj}}})$$
$$Z^{(L)} \in \mathbb{R}^{n \times 1}$$

The covariance matrix is constructed as formula (58).

$$K_{\text{learned}} = Z^{(L)} (Z^{(L)})^T \tag{58}$$

Bayesian uncertainty enhancement is defined as formula (59).

$$K_{\text{bayes}} = K_{\text{learned}} + \Sigma_{\text{uncertainty}} \tag{59}$$

Among them, the uncertainty matrix is $\Sigma_{\text{uncertainty}} = \text{diag}(\sigma^2_{\text{epistemic}}) + U_{\text{aleatoric}}$.

The parameter uncertainty propagation is expressed by formula (60)

$$\sigma^2_{\text{epistemic}}(x_i) = \sum_{l=1}^{L} \sum_{p \in \{\text{Q,K,V,O}\}} \text{Tr}[\frac{\partial H^{(l)}}{\partial W_p^{(l)}} \Sigma_{q,W_p^{(l)}} (\frac{\partial H^{(l)}}{\partial W_p^{(l)}})^T] \tag{60}$$

Heteroscedasticity modeling[43] for formula (61).

$$\sigma^2_{\text{aleatoric}}(x_i) = \exp(f_{\text{var}}(x_i)) \tag{61}$$

Among them, f_{var} is an independent variance prediction network, represented by formula (62).

$$f_{\text{var}}(x_i) = H_i^{(L)} W_{\text{var}} + b_{\text{var}} \tag{62}$$

The variance network parameter prior is obtained from Equations (81) to (63).

$$b_{\text{var}} \sim N(\mu_{b_{\text{var}}}, \Sigma_{b_{\text{var}}}) \tag{63}$$

$$W_{\text{var}} \sim N(\mu_{W_{\text{var}}}, \Sigma_{W_{\text{var}}}) \tag{64}$$

The joint uncertainty matrix was finally determined as formula (65).

$$U_{\text{aleatoric}} = \text{diag}(\sigma^2_{\text{aleatoric}}(x_1), \ldots, \sigma^2_{\text{aleatoric}}(x_n)) \tag{65}$$

Symmetrization treatment[44] is given by formula (66).

$$K_{\text{sym}} = \frac{1}{2}(K_{\text{bayes}} + K^T_{\text{bayes}}) \tag{66}$$

Eigenvalue decomposition[45] into formula (67).

$$K_{\text{sym}} = U\Lambda U^T \tag{67}$$

Positive correction is corrected to formula (68).

$$\tilde{\Lambda} = \max(\Lambda, \epsilon I) \tag{68}$$

Among them, $\epsilon > 0$ is the numerical stability parameter.The corrected covariance matrix is given by formula (69)

$$K_{\text{final}} = U\tilde{\Lambda}U^T \tag{69}$$

The task covariance matrix is expressed by formula (70).

$$K_{\text{task}} \in \mathbb{R}^{T\times T} \tag{70}$$

Among them, T is the number of tasks, K_{task} which obeys the inverse Wishart prior and can be expressed by formula (71).

$$K_{\text{task}} \sim \text{IW}(\Psi, \nu) \tag{71}$$

The complete multitask covariance is defined by the product structure in formula (72).

$$K_{\text{multi}} = K_{\text{final}} \otimes K_{\text{task}} \tag{72}$$

When the task is independent, use block diagonal approximation optimization, defined as formula (73).

$$K_{\text{multi}} = \text{blockdiag}(K^{(1)}_{\text{final}}, \ldots, K^{(T)}_{\text{final}}) \tag{73}$$

The covariance hyperparameter learning process is as follows.

The global scaling factor is defined by formula (74).

$$\tau^2 \sim \text{Gamma}(\alpha_\tau, \beta_\tau) \tag{74}$$

The final covariance matrix is defined by formula (75).

$$K = \tau^2 K_{\text{final}} + \sigma^2_n I \tag{75}$$

The covariance parameter is variational posterior updated to formula (76).

$$q(\tau^2) = \text{Gamma}(\tilde{\alpha}_\tau, \tilde{\beta}_\tau) \tag{76}$$

The update equation is defined by formulas (77) and (78).

$$\tilde{\alpha}_\tau = \alpha_\tau + \frac{n}{2} \tag{77}$$

$$\tilde{\beta}_\tau = \beta_\tau + \frac{1}{2}\text{Tr}[K^{-1}_{\text{final}}(y-\mu)(y-\mu)^T] \tag{78}$$

The process for assessing the quality of the covariance is as follows.

Condition number monitoring is performed using formula (79).

$$\kappa(K) = \frac{\lambda \text{max}(K)}{\lambda_{\text{min}}(K)} \tag{79}$$

Rank deficiency detection is given by formula (80).

$$\text{rank}(K) = \sum_{i=1}^{n} \mathbf{1}[\lambda_i > \epsilon_{\text{rank}}] \tag{80}$$

The Frobenius norm condition[46] is given by formula (81).

Attribute	Description
Data Scale	768 samples × 8 features
Target Variable	Y1 (Energy consumption indicator)
Feature Type	Continuous numerical and categorical
Main Feature Domain	Building parameters - Relative compactness (X1), Surface area (X2), Wall area (X3), Roof area (X4), Overall height (X5), Orientation (X6), Glazing area (X7), Glazing area distribution (X8)
Application Scenario	Building energy consumption prediction, green building design, energy-saving optimization
Data Characteristics	Building engineering data, predicting heating load based on building physical parameters, suitable for energy efficiency modeling and sustainable building design

Table 1. Energy efficiency Dataset

Attribute	Description
Data Scale	17,520 samples × 7 features + time index
Time Range	Hourly data from November 2008 onwards
Feature Type	Continuous numerical + timestamps
Main Feature Domain	Global active power (Global_active_power), Global reactive power (Global_reactive_power), Voltage (Voltage), Global current intensity (Global_intensity), Kitchen electricity consumption (Sub_metering_1), Laundry room electricity consumption (Sub_metering_2), Water heater and air conditioning electricity consumption (Sub_metering_3)
Application Scenario	Time series forecasting, energy consumption analysis, smart grid, load forecasting
Data Characteristics	Long-term continuous time series data containing seasonal and periodic patterns, suitable for temporal modeling and energy consumption prediction tasks

Table 2. Household_power_timeseries

$$K_F = \sqrt{\sum_{i,j} K_{ij}^2} \tag{81}$$

The logarithmic determinant of the covariance matrix is given by formula (82).

$$\log|K| = \sum_{i=1}^{n} \log(\lambda_i) \tag{82}$$

The Bayesian updating rule is as follows.

The likelihood function is given by formula (83).

$$p(y|X,K) = N(y;\mu,K) \tag{83}$$

The posterior covariance is given by formula (84).

$$K_{\text{post}} = (K^{-1} + \Phi^T \Sigma_y^{-1} \Phi)^{-1} \tag{84}$$

where Φ is the feature mapping matrix, and Σ_y is the observation noise covariance.The test point prediction covariance is given by formula (85).

$$\text{Cov}[f_*|X,y,X_*] = K_{**} - K_{*X} K_{XX}^{-1} K_{X*} \tag{85}$$

Among them: $K_{**} = K(X_*, X_*)$ is the covariance between test points, $K_{*X} = K(X_*, X)$ is the cross-covariance between test and training points, and$K_{XX} = K(X, X)$ is the covariance of training points.

Model experiments

This experiment conducts testing on six datasets, with specific dataset information referenced in Tables 1, 2, 3, 4, 5 and 6. The following metrics are primarily selected to evaluate experimental results (Eqs. 86–91): R^2 measures the proportion of target variable variance explained by the model, reflecting the overall accuracy of predictions.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{86}$$

Where y_i represents the true values (actual observed data), $\hat{y}_i$ represents the model predicted values, and$\bar{y}$ represents the mean of true values. The Prediction Interval Coverage Probability (PICP) metric is introduced

Attribute	Description
Data Scale	20,640 samples × 8 features
Target Variable	target (median house price, unit: hundred thousand dollars)
Feature Type	Continuous numerical + geographic coordinates
Main Feature Domain	Median income (MedInc), House age (HouseAge), Average rooms (AveRooms), Average bedrooms (AveBedrms), Population (Population), Average occupancy (AveOccup), Latitude (Latitude), Longitude (Longitude)
Geographic Range	Statewide California, based on 1990 census data
Application Scenario	House price prediction, geospatial modeling, regression analysis benchmarking
Data Characteristics	Contains geographic coordinate information, suitable for spatial correlation analysis and geographically weighted regression modeling

Table 3. California housing Dataset

Attribute	Description
Data Scale	395 samples × 15 features
Target Variable	G3 (final grade, 0–20 scale)
Feature Type	Numerical, ordinal
Main Feature Domain	Demographic characteristics (age), family background (parental education level), learning behavior (study time, absenteeism), social behavior (going out frequency, alcohol consumption), health status, historical academic performance (G1, G2)
Application Scenario	Student performance prediction, educational data mining
Data Source	Portuguese high school student survey data

Table 4. Student_Performance Dataset

Attribute	Description
Data Scale	768 samples × 8 features
Target Variable	Y1, Y2 (dual-output regression task)
Feature Type	Continuous numerical
Main Feature Domain	Process parameters (X1-X8), including relative compactness, surface area, wall thickness, silicon content and other building material or engineering parameters
Output Variables	Y1 and Y2 represent two performance indicators of the system
Application Scenario	Multi-output regression, engineering optimization, performance prediction
Data Characteristics	Engineering/manufacturing data, suitable for multi-objective optimization and process parameter modeling

Table 5. Power_Plant Dataset

Attribute	Description
Data Scale	308 samples × 6 features
Target Variable	Residuary_Resistance (residual resistance)
Feature Type	Continuous numerical
Main Feature Domain	Hull geometric parameters - Longitudinal center of buoyancy position (LongPos_COB), Prismatic coefficient (Prismatic_Coeff), Length displacement ratio (Len_Disp_Ratio), Beam draft ratio (Beam_Draft_Ratio), Length beam ratio (Length_Beam_Ratio), Froude number (Froude_Num)
Application Scenario	Ship design optimization, fluid dynamics modeling, resistance prediction
Data Characteristics	Marine engineering data, predicting hydrodynamic performance based on different hull design parameters, suitable for engineering regression and optimization problems

Table 6. Yacht_Hydrodynamics Dataset

to evaluate the reliability of prediction intervals[47], which reflects the probability that actual observed values fall within the upper and lower bounds of the prediction intervals.

$$\mathrm{PICP} = \frac{1}{N_t}\sum_{i=1}^{N_t}\kappa_i^{(\alpha)} \tag{87}$$

Where N_t is the number of prediction samples, κ is a boolean variable, and α is the confidence level parameter. The Prediction Interval Normalized Average Width (PINAW) metric is introduced to reflect the conservativeness of predictions[48–50]. Conservativeness due to purely pursuing reliability results in prediction intervals that are excessively wide, failing to provide effective uncertainty information for predicted values and losing decision-making value.

$$\mathrm{PINAW} = \frac{1}{N_t R}\sum_{i=1}^{N_t}\left[\tilde{U}_i^{(\alpha)}(x_i) - \tilde{L}_i^{(\alpha)}(x_i)\right] \tag{88}$$

Where N_t is the number of prediction samples, *R* is the range of prediction target values used for normalizing the average width, $\tilde{U}_i^{(\alpha)}(x_i)$ is the upper bound of the prediction interval for the *th* sample, $\tilde{L}_i^{(\alpha)}(x_i)$ is the lower bound of the prediction interval for the *i-th* sample, and α is the confidence level parameter. Expected Calibration Error (ECE) is an important metric for measuring model calibration performance[51,52], used to evaluate the degree of consistency between the model's predicted confidence and actual accuracy[9].

$$\mathrm{ECE} = \sum_{b=1}^{B}\frac{n_b}{N}\left|\mathrm{acc}(b) - \mathrm{conf}(b)\right| \tag{89}$$

Where bin refers to dividing the continuous confidence range into several discrete intervals, B represents the total number of bins, n_b represents the total number of samples in the *b-th* bin, N represents the total number of all samples, acc(b) represents the average of true labels for samples in the *b-th* bin, and conf(b) represents the average of model predicted confidence for samples in the *b-th* bin. Continuous Ranked Probability Score (CRPS) is an important metric for evaluating the quality of probabilistic predictions[53].

$$\mathrm{CRPS}(F, y) = \int \left(F(x) - \mathbf{1}_{\{x \geqslant y\}}\right)^2 dx \tag{90}$$

Where $F(x)$ represents the cumulative distribution function (CDF) predicted by the model[54], which describes the probability that a random variable is less than or equal to a certain value *x*. *y* is the observed true value, i.e., the actual outcome that occurred. *x* is the integration variable, representing all possible value ranges. The indicator function $\mathbf{1}\{x \geqslant y\}$ is a step function: when *x*. is greater than or equal to the true value *y*, the function value is 1; when *x*. is less than *y*, the function value is 0.

Area Under the Sparsification Error curve (AUSE) measures the degree of consistency between uncertainty scores and true errors, i.e., the extent to which uncertainty estimates can reflect the model's true mistakes[55].

TorchUncertainty has not been formally defined by the official source, but based on its definition, it can be expressed as formula (91).

$$\mathrm{AUSE} = \int_0^1 \mathrm{SE}(f)\,df \tag{91}$$

Where $\mathrm{SE}(f)$ is the sparsification error after removing a fraction *f* of high uncertainty samples, and $f \in [0, 1]$ represents the proportion of removed samples.

This research introduces Deep Ensemble, Monte Carlo Dropout (MC Dropout), Bayesian Deep Learning (BDL), and Gaussian Process (GP) methods for comparison with RBA, with conceptual explanations as follows.

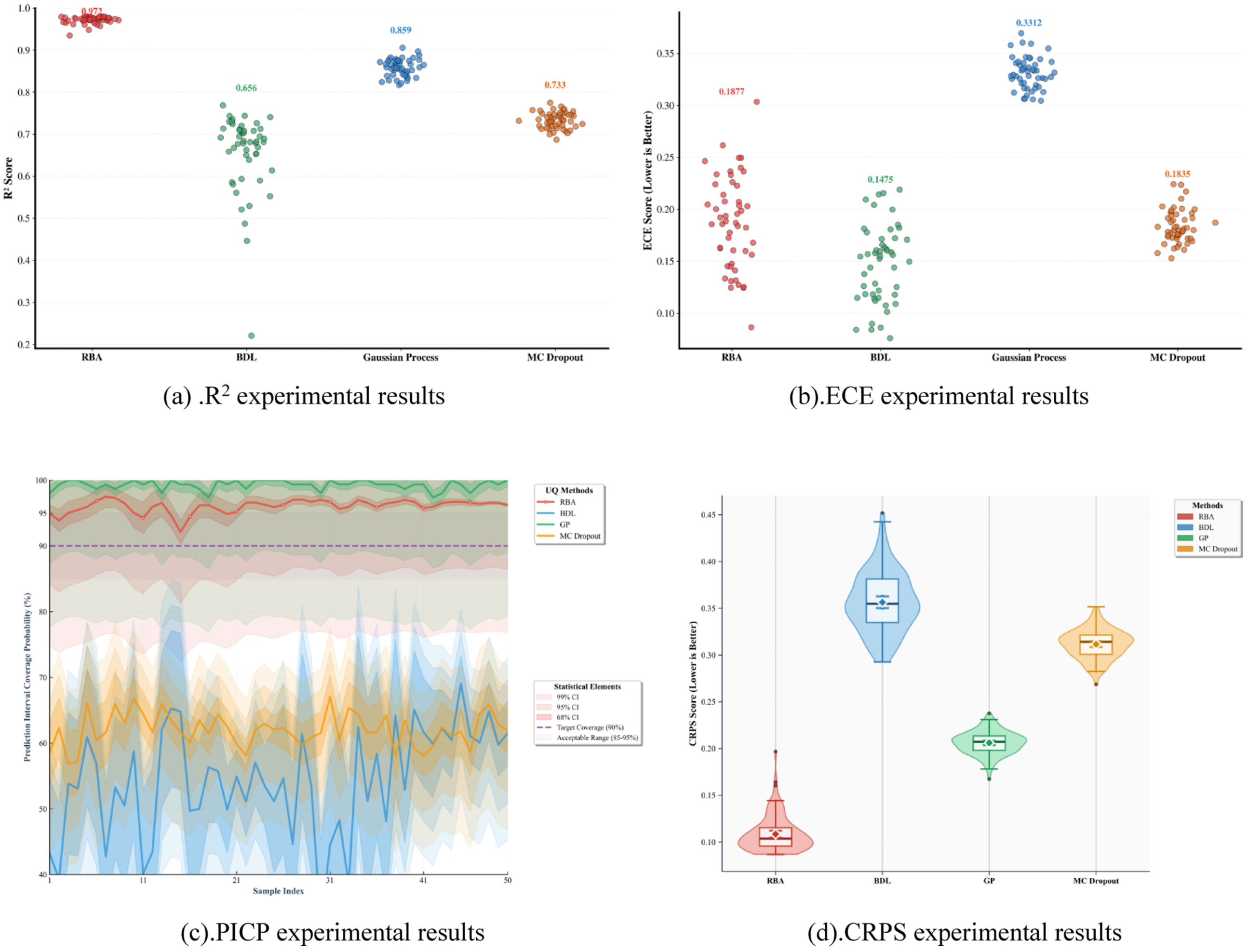
The Deep Ensemble method is based on statistical ensemble theory, constructing predictive ensemble models by training multiple independently initialized deep neural networks. This method follows the theoretical framework of the Condorcet jury theorem, utilizing weight distribution differences and loss trajectory diversity among models to achieve marginal likelihood estimation of predictive distributions. In practice, this method employs negative log-likelihood as the loss function, enabling individual networks to simultaneously predict mean and variance, obtaining well-calibrated uncertainty estimates through ensemble inference.

Monte Carlo Dropout is built on the theoretical foundation of variational inference, reinterpreting dropout regularization as approximate Bayesian inference in deep Gaussian processes. This method achieves Monte Carlo sampling approximation of weight posterior distributions by maintaining dropout activation during the inference phase. Theoretically, this technique transforms deterministic neural networks into stochastic models, generating predictive distributions through multiple forward passes to quantify the model's epistemic uncertainty. Its theoretical advantage lies in achieving Bayesian neural network approximation without modifying existing network architectures, though the quality of its uncertainty estimates is significantly affected by dropout rates and weight regularization parameters.

The Bayesian Deep Learning (BDL) framework combines Bayesian statistical principles with deep learning architectures, achieving probabilistic representation of parameters by imposing prior distributions on neural network weights and updating posterior distributions using observed data. This method can distinguish and quantify two types of uncertainty: aleatoric uncertainty (reflecting the inherent randomness of observation noise) and epistemic uncertainty (reflecting insufficient knowledge of model parameters). The theoretical foundation of BDL stems from Bayes' theorem, approximating intractable posterior distributions through variational inference or Markov Chain Monte Carlo methods, providing a principled uncertainty quantification framework for safety-critical applications.

Gaussian Process (GP) constitutes a non-parametric Bayesian method that defines probability distributions over function spaces to achieve regression and classification tasks. The theoretical foundation of GP is built on random process theory, where the joint distribution of any finite number of function values follows a multivariate Gaussian distribution. This method encodes prior knowledge through kernel functions and obtains posterior predictive distributions by combining observed data using the Bayesian inference framework. The core advantage of GP lies in its analytical tractability, directly providing closed-form solutions for predictive mean and variance without requiring additional uncertainty quantification steps. In uncertainty quantification applications, GP is widely used for surrogate modeling, Bayesian optimization, and sensitivity analysis tasks.

This study first uses a lower-dimensional dataset Energy Efficiency, which belongs to small data and contains engineering features. As shown in the experimental results in Fig. 2: (a) RBA achieved a prediction accuracy of approximately 0.972, surpassing the other three methods, including Bayesian Deep Learning (BDL) at 0.656, Gaussian Process at 0.859, and MC Dropout at 0.741. This performance difference is further validated in the analysis in (b), where RBA achieved a relatively low calibration error of 0.1877. The temporal analysis in (c) demonstrates the dynamic behavioral characteristics of each method in uncertainty quantification. RBA and Gaussian Process maintained close to or exceeding 90% target coverage probability at most sampling points, demonstrating good uncertainty calibration capability. In the evaluation in (d), RBA's distribution concentrates in lower numerical intervals, indicating that its predicted probability distributions have higher consistency with true observed values. In contrast, BDL and MC Dropout show higher CRPS values, meaning their probabilistic predictions exhibit greater deviations. Although Gaussian Process performs stably in certain cases, the variability in its CRPS distribution indicates its limitations in handling complex data structures.



(a) .R^2 experimental results

(b).ECE experimental results

(c).PICP experimental results

(d).CRPS experimental results

Fig. 2. Energy Efficiency experimental results. (**c**).PICP experimental results, (**d**).CRPS experimental results

To validate RBA's generalization capability, this study further introduces the larger-scale time series prediction regression task dataset in Table 2 to explore the dataset characteristics that RBA optimally adapts to. Figure 3 shows that the RBA model has a coefficient of determination of 0.920, root mean square error of 0.23 kW, and correlation coefficient of 0.963. From the prediction-actual value scatter plot observation, RBA's data points are distributed around the ideal prediction line.

Figure 4 shows that RBA's residual mean is 0.030 kW, standard deviation is 0.236 kW, and trend line slope is −0.029, indicating relatively small systematic bias between residuals and predicted values. The BDL model has an uncertainty coverage probability of 0.948, but relatively low prediction accuracy with R² of 0.820 and RMSE of 0.35 kW, and the residual distribution exhibits heteroscedastic characteristics. Gaussian Process has R² of 0.902, RMSE of 0.26 kW, and CRPS score of 0.096. The MC Dropout method has R² of 0.842 and uncertainty coverage probability of 0.873, showing relatively weaker performance across multiple evaluation metrics.

Figure 5 shows that the RBA model performs stably in capturing power consumption variation patterns, with uncertainty bandwidth indicator PINAW of 0.180. When handling power peaks, RBA maintains relatively small prediction errors and moderate uncertainty ranges. Different models show variations in prediction performance across different time periods, with RBA demonstrating relatively consistent performance in maintaining the balance between prediction accuracy and uncertainty quantification. Experimental results indicate that RBA's residual attention mechanism has a certain role in improving prediction performance and uncertainty estimation.

Based on the Table 3 dataset, this study conducted both parallel and ablation experiments to explore the contribution of each RBA component. Figure 6 shows point estimation accuracy, where the RBA method has a root mean square error of 0.5089 and coefficient of determination of 0.8027; the MC Dropout method has corresponding indicators of 0.5131 and 0.7994 respectively; the BDL method has 0.5179 and 0.7957. In

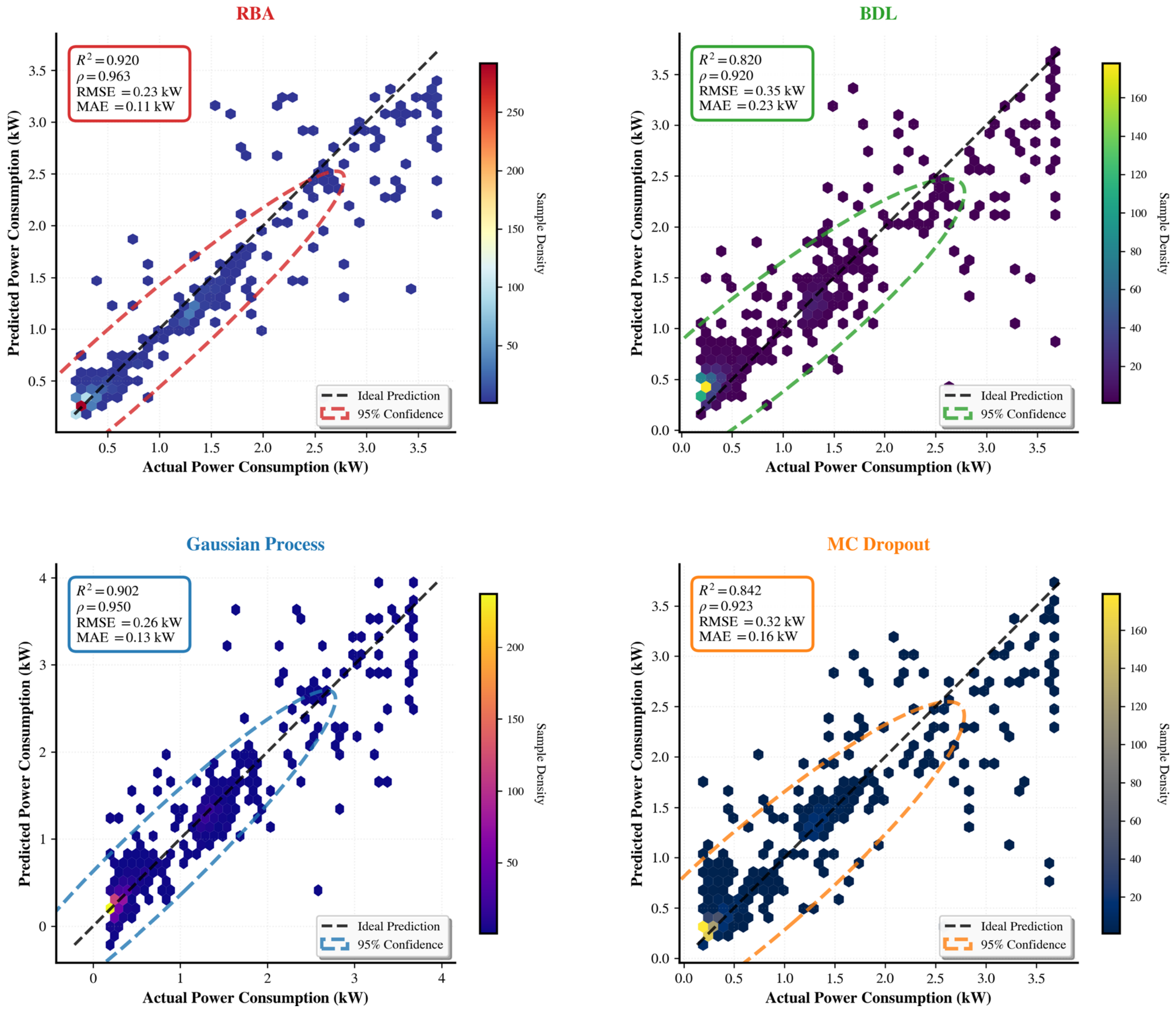


Fig. 3. Household_power_timeseries prediction accuracy experiment results

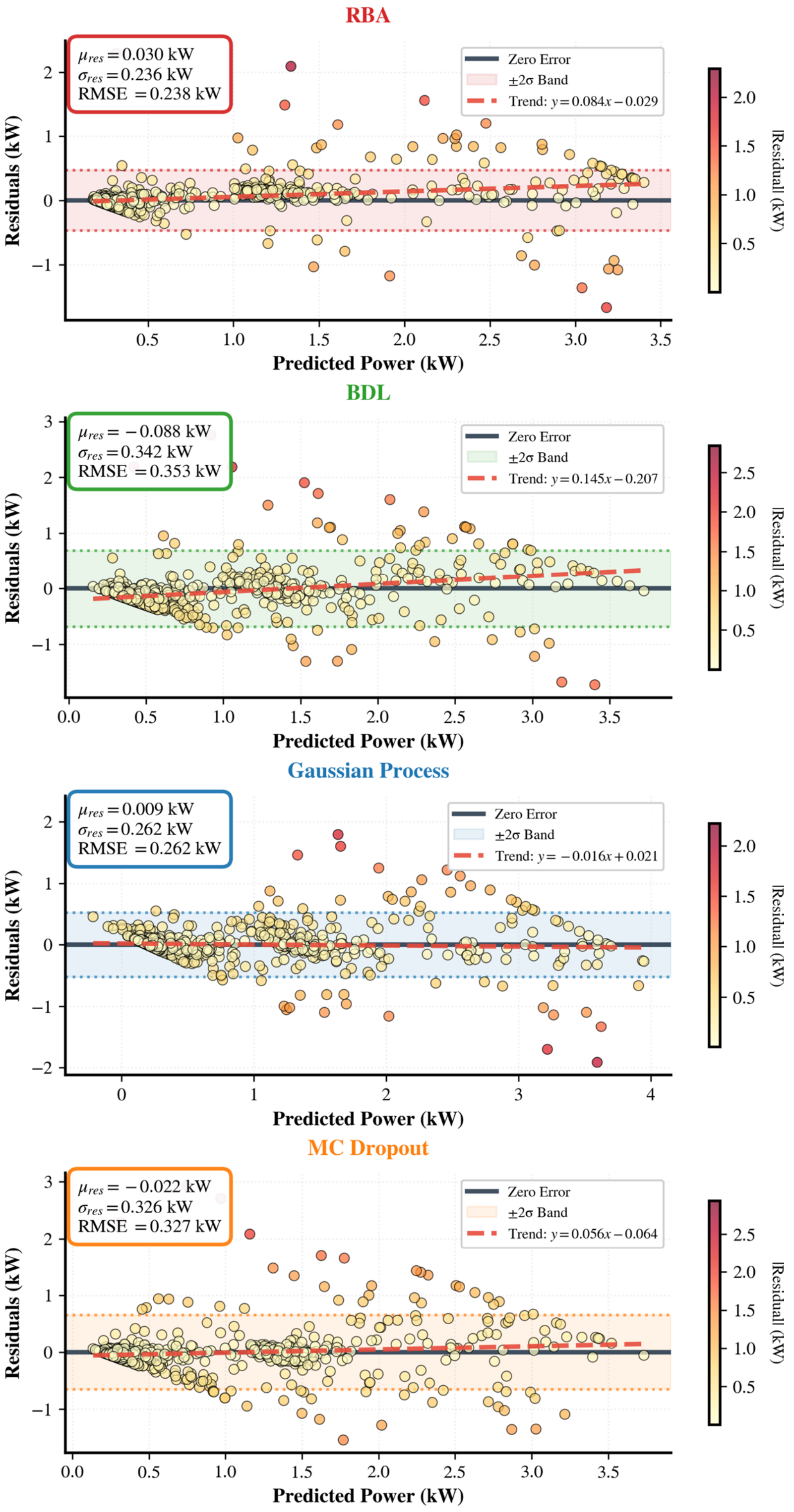


Fig. 4. Household_power_timeseries results of residual analysis experiment

geospatial visualization, all three methods can capture the basic spatial distribution characteristics of real house price data, with prediction results showing similarity to true values in spatial patterns.

Figure 7 shows uncertainty calibration evaluation, where the RBA method has an Expected Calibration Error of 0.0660, Prediction Interval Coverage Probability reaching 0.9638, and Continuous Ranked Probability Score of 0.2898. From spatial distribution characteristics observation, RBA method's uncertainty estimation presents specific patterns in geographic distribution, with high uncertainty regions mainly distributed in geographic locations with sparse data or dramatic house price changes.

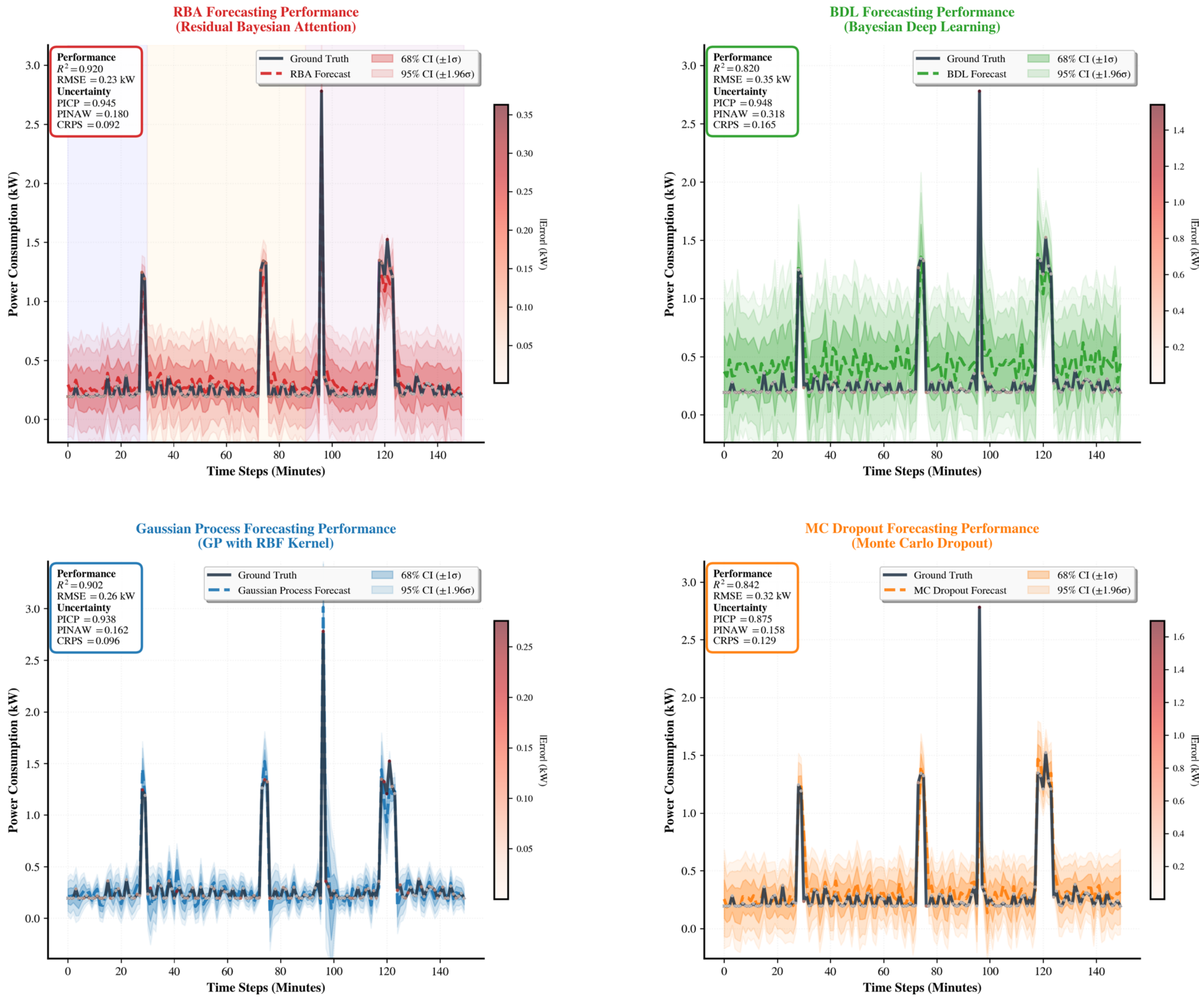


Fig. 5. Household_power_timeseries time series forecasting experiment results

Figure 8 shows performance differences among models in handling spatial dependency. Introducing Moran's I spatial autocorrelation analysis can quantitatively evaluate each model's capability to capture spatial structure. The RBA model's Global Moran's I value is 0.2211 with Z-Score of 37.250, indicating that its residuals still exhibit moderate spatial clustering. The MC Dropout model's Moran's I value rises to 0.2519 with Z-Score of 42.438, showing stronger spatial autocorrelation, meaning this model systematically overestimates or underestimates house prices in certain geographic regions. The BDL model performs worst, with Moran's I value reaching 0.2728 and Z-Score of 45.963, indicating its residuals have the strongest spatial clustering characteristics. From spatial distribution patterns, all three models show significant positive spatial autocorrelation (red regions) in the San Francisco Bay Area (longitude − 122° to −121°, latitude 37.5° to 38.5°) and Los Angeles area (longitude − 118° to −117°, latitude 33.5° to 34.5°), indicating that prediction errors in these high house price regions exhibit clustered distribution.

Based on the ablation experiment results analysis using the California housing dataset (Tables 7 and 8), experimental data reveals differential impacts of different Bayesian components on model performance. From regression performance indicators, the configuration removing the Bayesian feedforward network shows relative advantages across three core indicators, with RMSE of 0.6204, MAE of 0.4306, and.

R^2 reaching 0.7062. In contrast, the configuration removing the Bayesian attention mechanism shows obvious performance degradation, with RMSE rising to 0.6516 and R^2 declining to 0.6760. The configuration removing the Bayesian covariance component lies between the two, showing moderate performance.

Cross-validation results further supplement the evaluation dimension of model stability. From the coefficient of variation perspective, the configuration removing Bayesian covariance exhibits higher stability with CV coefficient of 0.0111, while the configuration removing Bayesian attention shows relatively larger performance fluctuations with CV coefficient reaching 0.0207. This difference indicates significant distinctions in how different components affect model generalization capability.

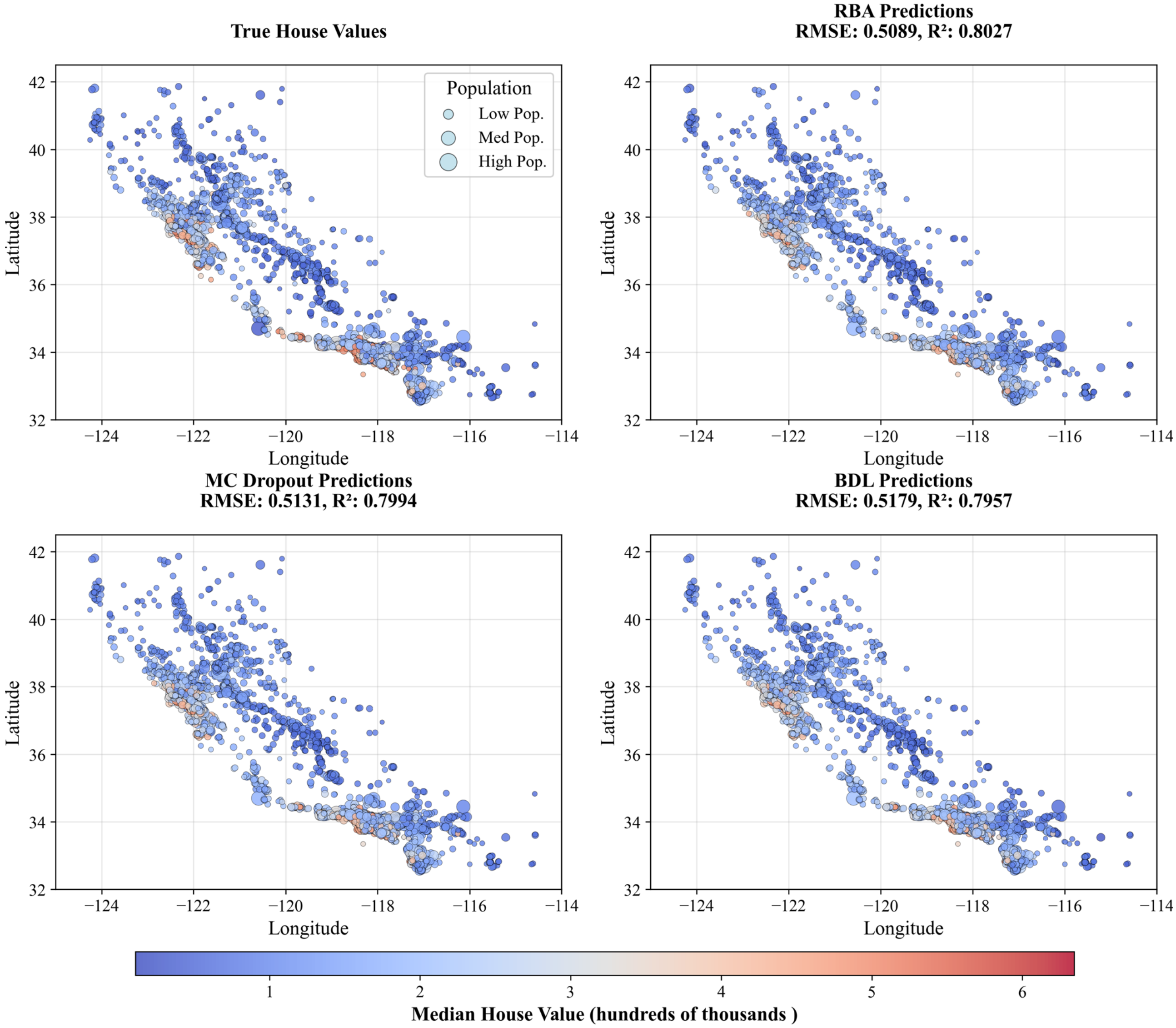


Fig. 6. California Housing Dataset Experimental eesults on spatial mapping relationships

Experimental results indicate that the Bayesian attention mechanism plays an important functional role in the model architecture, with its absence leading to dual degradation in both performance and stability. The Bayesian covariance component primarily contributes to the stability dimension, having a positive effect on the model's cross-dataset generalization.

To improve the universality of experimental results, this study introduces the Deep Ensemble algorithm to repeat experiments on new datasets, conducting in-depth experimental testing using three models (MC Dropout, Deep Ensemble, RBA) on two strong nonlinear datasets (Student Performance Dataset, Power Plant Dataset) and one nonlinear dataset with physical property coupling (Yacht Hydrodynamics Dataset). Each dataset is divided according to an 80% training, 16% validation, 4% testing ratio. Data preprocessing adopts standardized scaling, ensuring features and target variables have zero mean and unit variance, eliminating the impact of dimensional differences on model training. This experiment is based on the PyTorch framework, accelerated using CUDA 11.8.

The experimental results in Fig. 9 are as follows: (a) The R^2 performance plot reveals dramatic fluctuation characteristics among the three methods, with the RBA method showing substantial variation from 61.5% to 73.8%, while MC Dropout and Deep Ensemble oscillate in ranges of 59.8%–67.2% and 58.3%–71.4% respectively. This high inter-trial variability reflects the inherent complexity and nonlinear characteristics of student performance data. However, in terms of prediction interval coverage probability, the RBA method maintains stable coverage of 85.3%–95.2%, approaching the theoretical target value of 95%, while MC Dropout only reaches a suboptimal level of 69.8%–78.5%, and Deep Ensemble shows good calibration at 89.7%–94.1%. Meanwhile, the PINAW indicator shows that RBA achieves a balance between coverage and precision within an interval width of 41.2%–51.3%, superior to Deep Ensemble's range of 47.8%–56.2%.

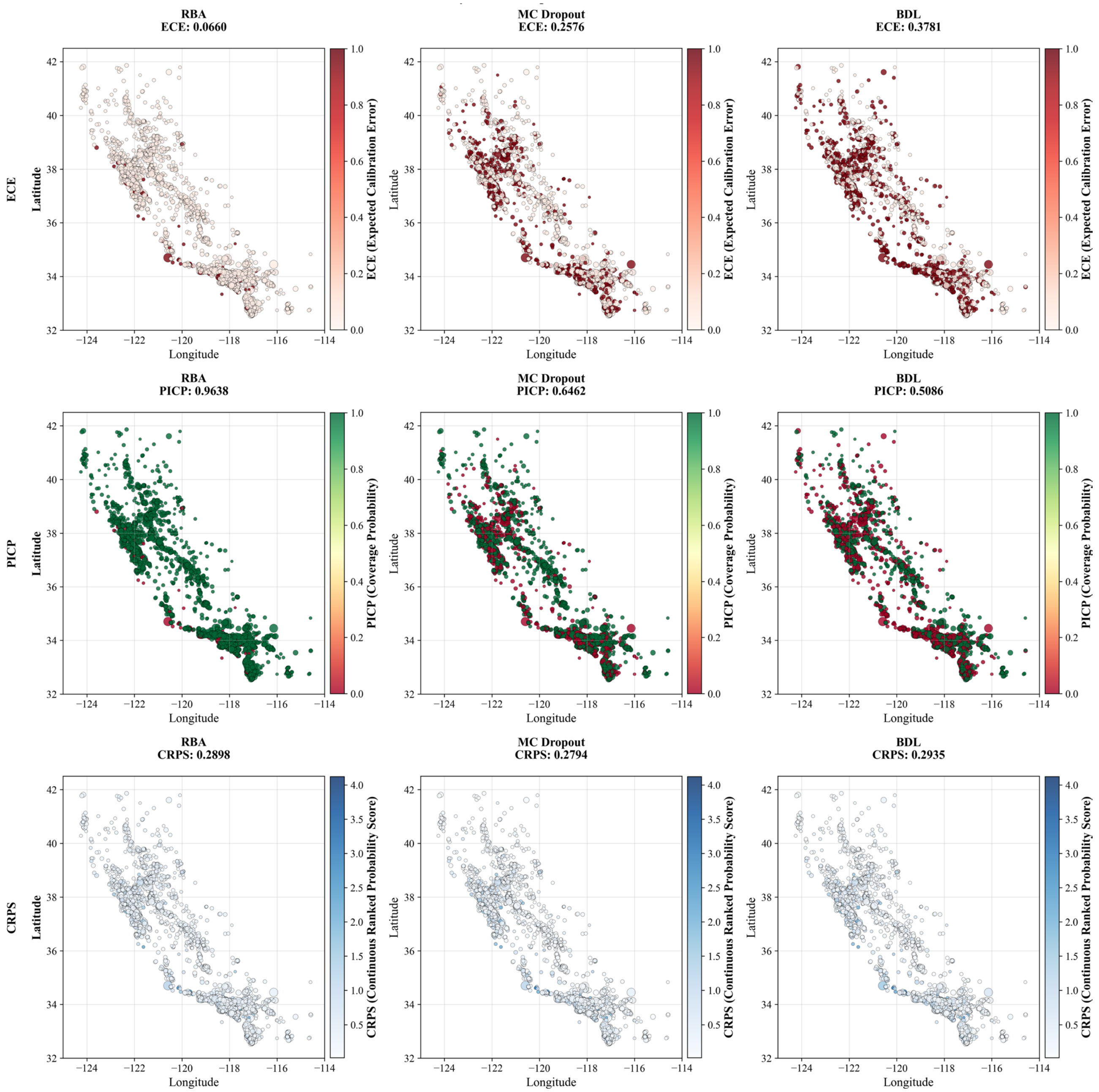
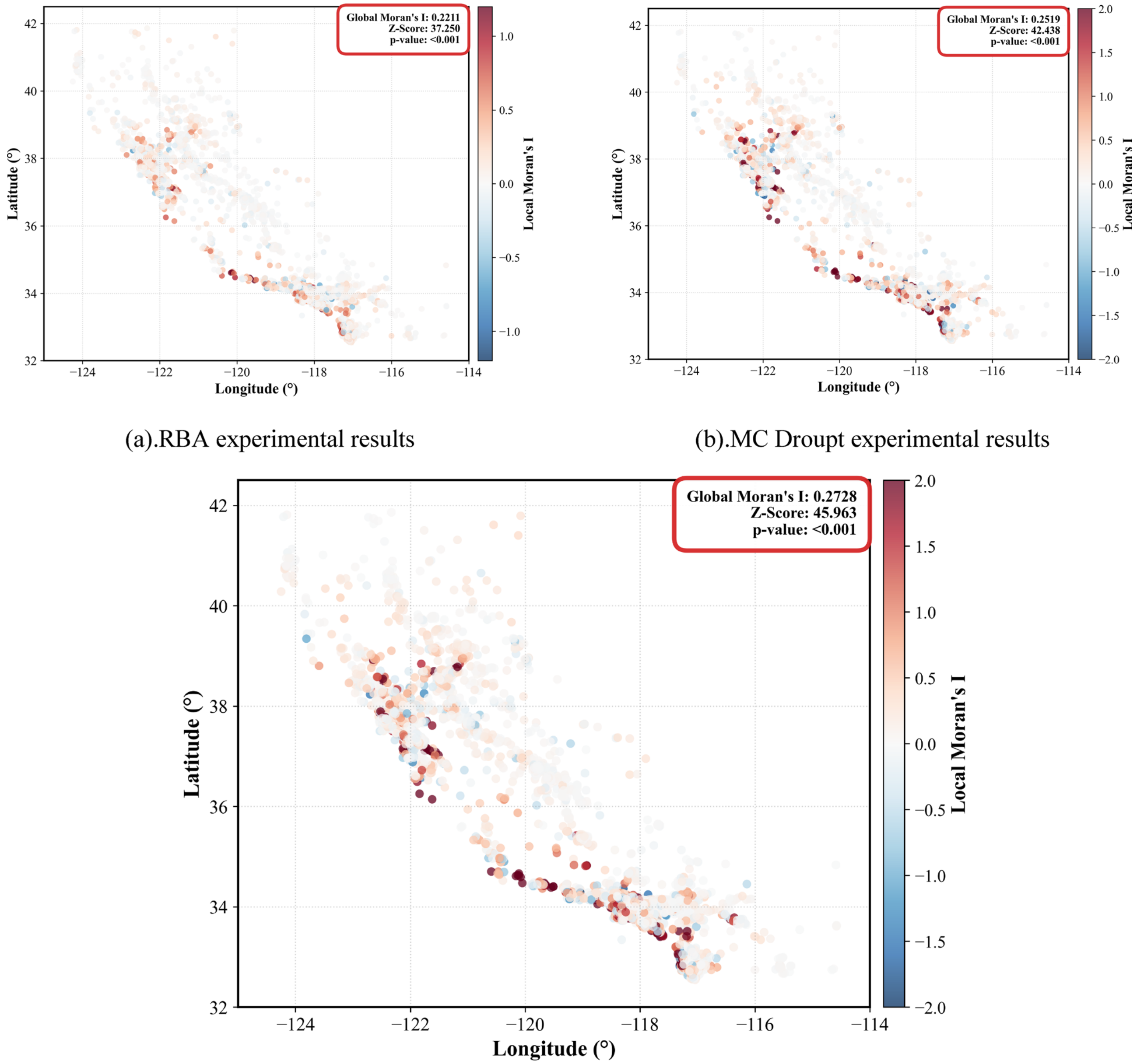


Fig. 7. California Housing Dataset experimental results of uncertainty analysis

a. (b) Shows distinctly different performance characteristics, with extreme volatility in R^2 values being particularly prominent. Each method oscillates dramatically between peaks of 84.8%–85.4% and valleys of 84.3%–84.9%. This "sawtooth" pattern suggests possible periodic disturbances or systematic noise in power plant operational data. Both RBA and Deep Ensemble maintain excellent coverage rates above 95% on the PICP indicator, while MC Dropout remains at an insufficient level of 88.7%–91.2%, indicating systematic underfitting problems when processing industrial process data. PINAW results further confirm that the RBA method achieves the narrowest prediction intervals within the range of 22.1%–27.8%, showing approximately 15% efficiency improvement compared to Deep Ensemble's 23.4%–29.7%.
b. (c) Presents the most extreme performance differentiation, with R^2 indicators showing negative value phenomena. The RBA method fluctuates dramatically between − 27.3% and 3.2%, MC Dropout maintains at −22.1% to 1.8%, while Deep Ensemble shows persistent negative values in the range of −18.5% to −14.2%. This widespread negative R^2 phenomenon indicates that complex fluid-structure interactions in yacht resistance prediction tasks exceed the expressive capabilities of current model architectures. However, paradoxically, Deep Ensemble shows the highest PICP coverage of 82.1%–98.3% in uncertainty quantification indicators, but this high coverage is achieved at the cost of extremely wide PINAW of 60.3%–87.4%, indicating that it compensates for severe prediction accuracy deficiencies through overly conservative interval estimation.



(a).RBA experimental results

(b).MC Droupt experimental results

(c).BDL experimental results

Fig. 8. California Housing Dataset results of the spatially dependent experiment. (**a**).RBA experimental results, (**b**).MC Droupt experimental results, (**c**).BDL experimental results

Model Configuration	RMSE	MAE	R^2
w/o Bayesian Attention	0.6516	0.4604	0.6760
w/o Bayesian FF	0.6204	0.4306	0.7062
w/o Bayesian Covariance	0.6351	0.4465	0.6921

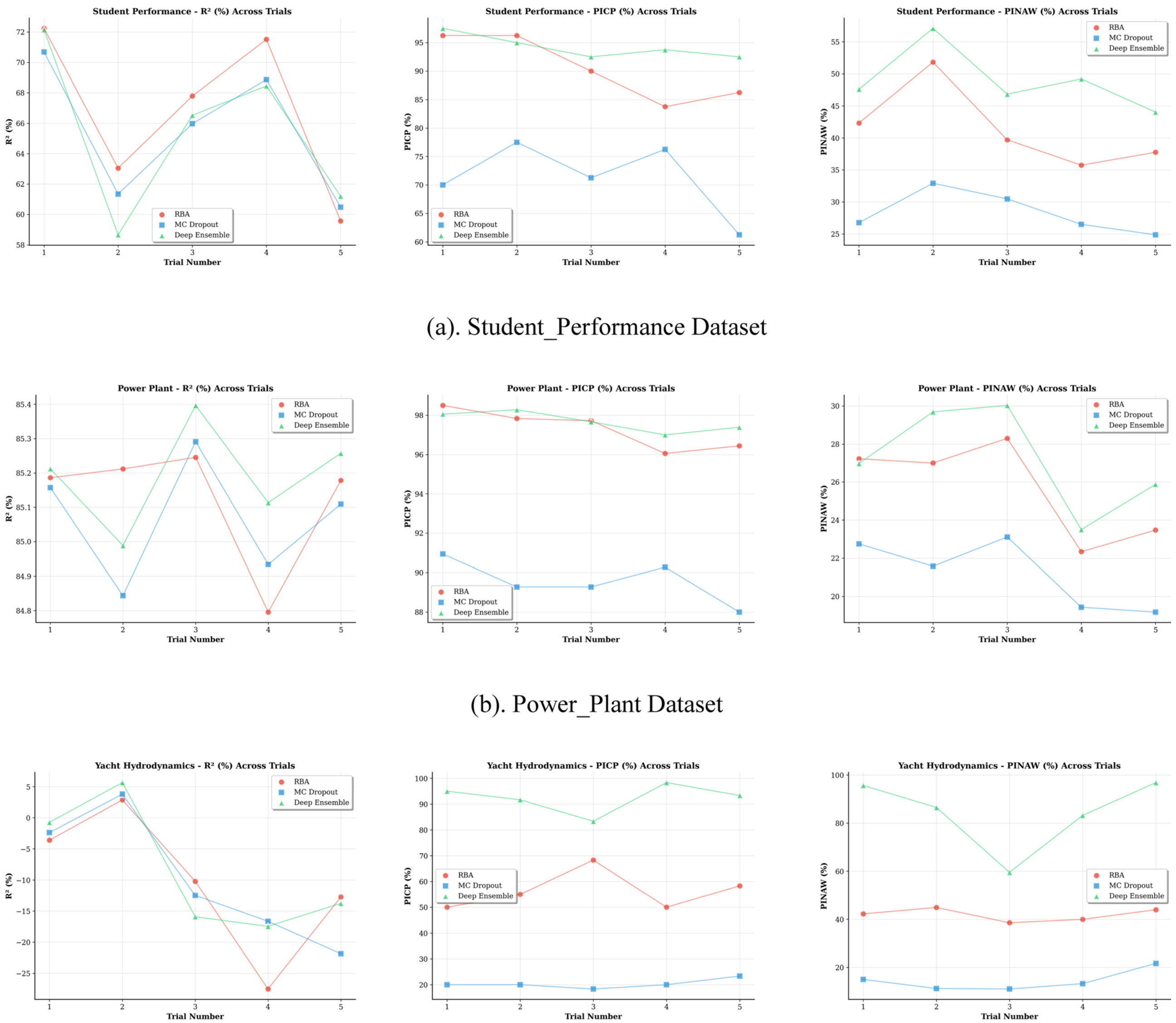
Table 7. Core regression performance metrics

In contrast, RBA controls PINAW within a narrower range of 40.1%–43.2% while maintaining reasonable coverage, demonstrating a more balanced uncertainty quantification strategy.

Figure 10 illustrates the reasons for the generally poor performance of models on the Yacht_Hydrodynamics Dataset. As the Froude number increases, the flow field undergoes a complete physical transition from viscosity-dominated displacement flow (Fr = 0.16) to a viscous-wave coupling transitional state (Fr = 0.32), and then to

Model Configuration	CV Mean RMSE	CV Std Dev	CV Coefficient
w/o Bayesian Attention	0.9181	0.0190	0.0207
w/o Bayesian FF	0.8542	0.0118	0.0138
w/o Bayesian Covariance	0.8781	0.0097	0.0111

Table 8. Cross -Validation robustness analysis



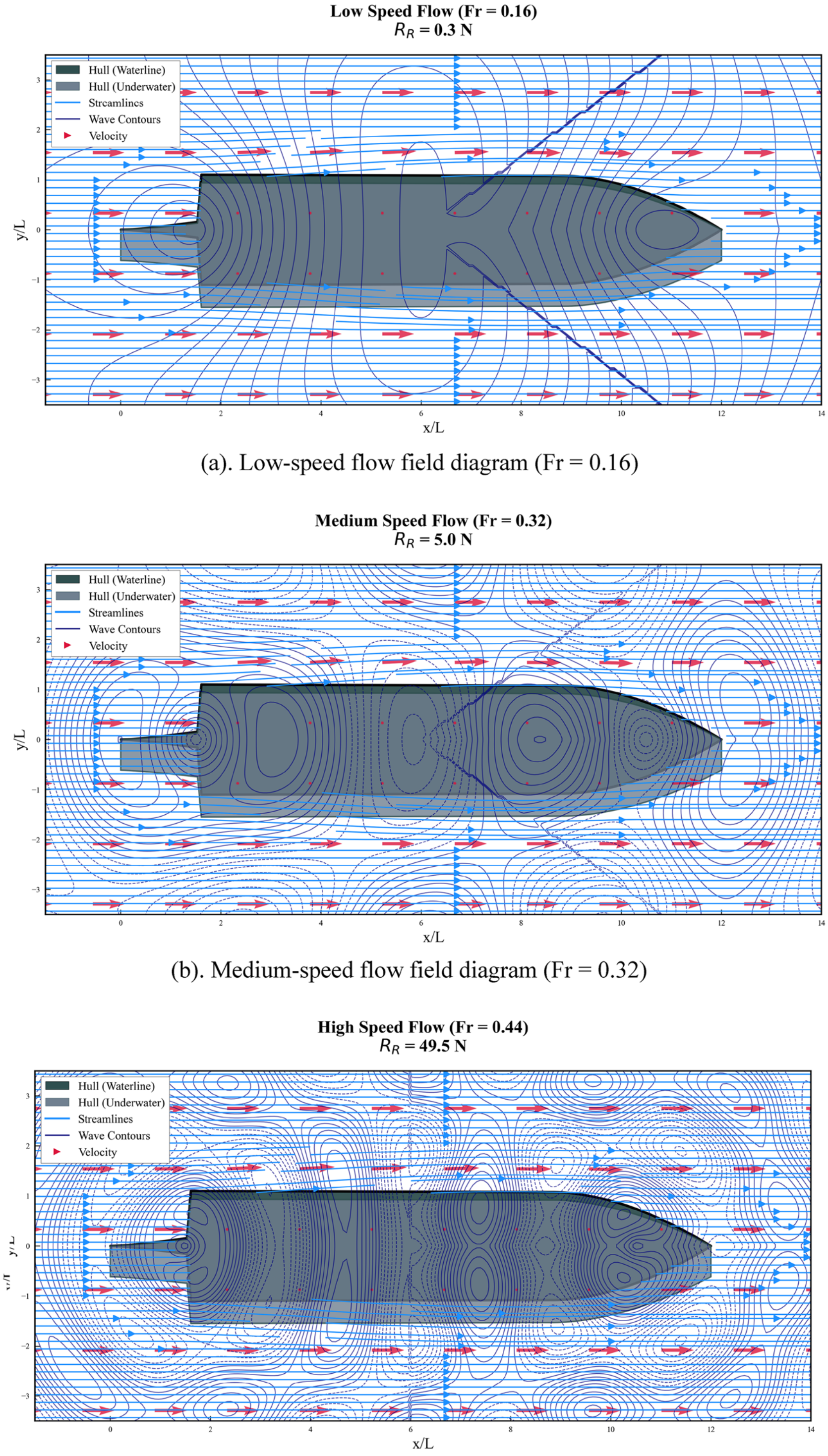
(a). Student_Performance Dataset

(b). Power_Plant Dataset

(c). Yacht_Hydrodynamics Dataset

Fig. 9. Deep validation of experimental results. (**a**). Student_Performance Dataset, (**b**). Power_Plant Dataset, (**c**). Yacht_Hydrodynamics Dataset

a wave resistance-dominated semi-planing regime (Fr = 0.44). Each stage corresponds to different governing equations and boundary conditions: at low speeds, it follows the Navier-Stokes equations for viscous flow; at high speeds, it requires consideration of potential flow theory under nonlinear free surface boundary conditions; while the transition region involves complex phenomena such as turbulent transition and flow separation. This multi-scale, multi-physics strongly coupled nonlinear characteristic manifests mathematically as discontinuous mapping relationships in high-dimensional parameter space, which fundamentally contradicts the smooth mapping mechanisms assumed by neural network architectures. This leads to inherent difficulties such as gradient vanishing, local optima, and limited generalization capability when RBA, Monte Carlo Dropout,



(a). Low-speed flow field diagram (Fr = 0.16)

(b). Medium-speed flow field diagram (Fr = 0.32)

(c). High-speed flow field diagram (Fr = 0.44)

Fig. 10. Evolution of Hydrodynamic Flow Fields in Yachts. (**a**). Low-speed flow field diagram, (Fr = 0.16), (**b**). Medium-speed flow field diagram, (Fr = 0.32), (**c**). High-speed flow field diagram, (Fr = 0.44)

and Deep Ensemble methods handle such marine hydrodynamic problems with distinct physical mechanism transitions.

The complexity of these physical mechanisms is further confirmed in Fig. 11, where all six key hull design parameters exhibit highly nonlinear relationships with residual resistance. Forward positioning of the longitudinal center of buoyancy triggers wave interference patterns leading to exponential resistance growth; the optimal slenderness ratio of the prismatic coefficient around 0.565 reflects the sensitive balance between pressure distribution and flow separation; the length-displacement ratio and beam-draft ratio embody the competing effects between hull fineness and stability requirements; while the Froude number exhibits the inherent cubic scaling characteristics of wave-making physics, establishing its status as the dominant parameter in resistance prediction. These complex geometric-hydrodynamic coupling relationships cannot be effectively captured through traditional machine learning's linear or simple nonlinear mapping, essentially requiring physics-informed modeling methods that incorporate naval architecture principles to accurately describe the multi-parameter interaction mechanisms in yacht resistance prediction.

Figure 12 presents the Froude number complexity analysis, clearly identifying three distinctly different operational domains: the viscous force-dominated displacement regime, the mixed physical characteristics semi-

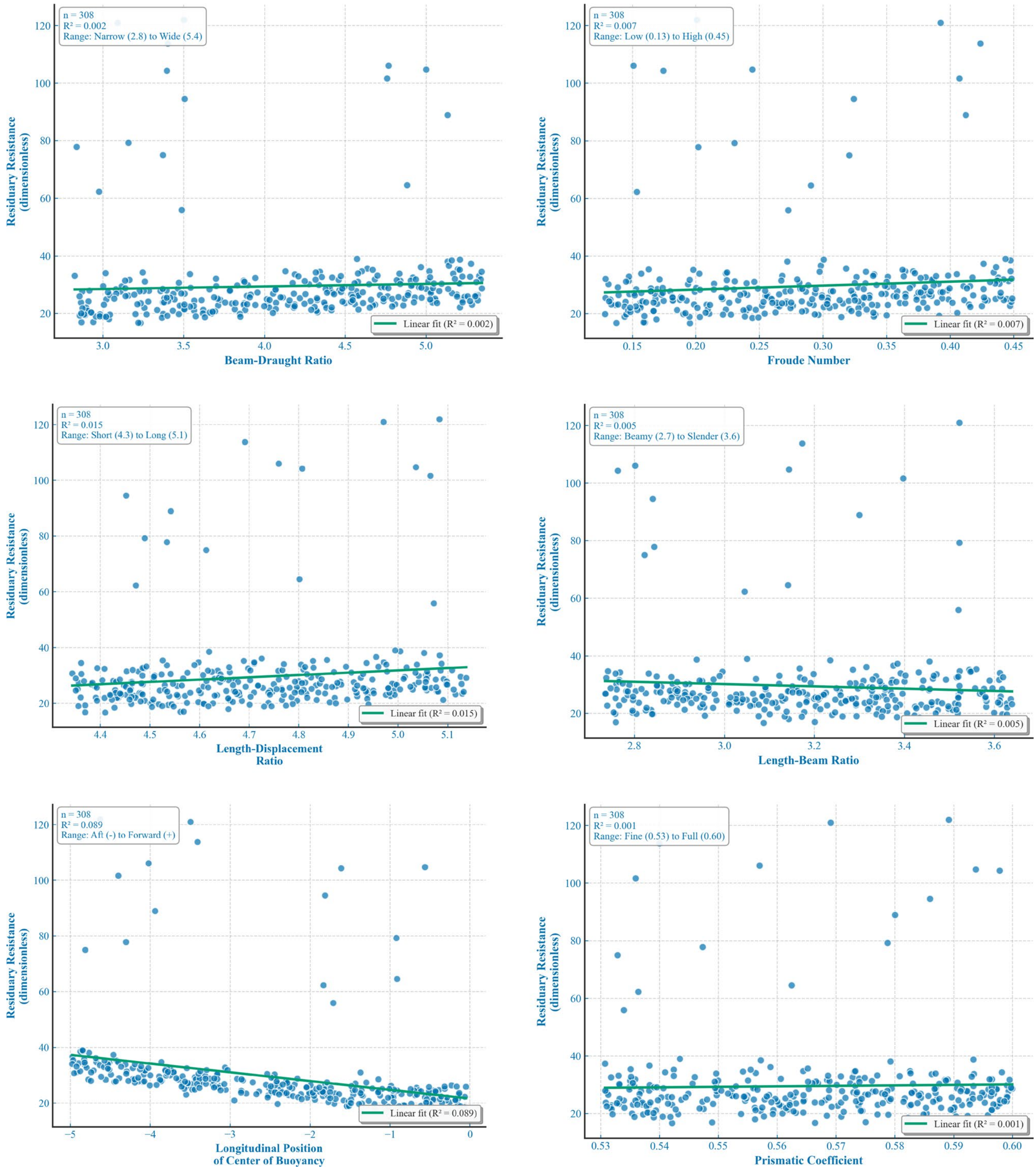


Fig. 11. Feature-Target relationship analysis

displacement transition region, and the planing regime where wave-making resistance exhibits dramatic cubic scaling. The discontinuous behavioral transitions that occur when hull speed crosses critical thresholds violate the continuity assumptions of standard statistical modeling. Residual analysis shows that although incorporating physical scaling laws achieves significant improvements, the complex interactions between hull shape factors, dynamic trim effects, and Reynolds number dependencies across operational regimes still generate substantial unexplained variance.

Based on the analysis of Figs. 10, 11, 12 and 50 independent repeatability experiments were conducted to primarily compare the lateral differentiation characteristics of the models. As shown in Fig. 13, (a) indicates that RBA and MC Dropout achieved R^2 values of 11.47%±14.24% and 9.66%±14.73%, respectively. Although the absolute values are not high, considering the physical complexity of yacht resistance prediction, these results are still within an acceptable range. However, the Deep Ensemble method exhibited extremely anomalous negative R^2 values (−359.88%±141.45%), a phenomenon indicating that this method experienced severe model failure when handling the specific physical constraints and nonlinear characteristics of yacht hydrodynamic data.

Figure 13 (b) shows that RBA and MC Dropout methods achieved ECE values of 0.441 ± 0.069 and 0.427 ± 0.058, respectively, which are statistically equivalent, while the Deep Ensemble method's ECE value reached 0.667 ± 0.051, significantly higher than the former two, indicating that this method suffers from a mismatch between prediction confidence and actual accuracy in modeling complex nonlinear systems such as yacht hydrodynamics.

Figure 13 (c) demonstrates that RBA AUSE values are primarily concentrated in the range of 5.65 ± 1.28, the Monte Carlo Dropout method achieved a mean of 5.97 ± 1.38, while the Deep Ensemble method's AUSE value reached as high as 25.35 ± 6.20, indicating systematic overestimation problems in its uncertainty estimation for this specific hydrodynamic task.

Discussion

From the systematic analysis of experimental results, a performance gradient phenomenon of profound significance can be observed: RBA's excellent performance on structured engineering data (Energy Efficiency

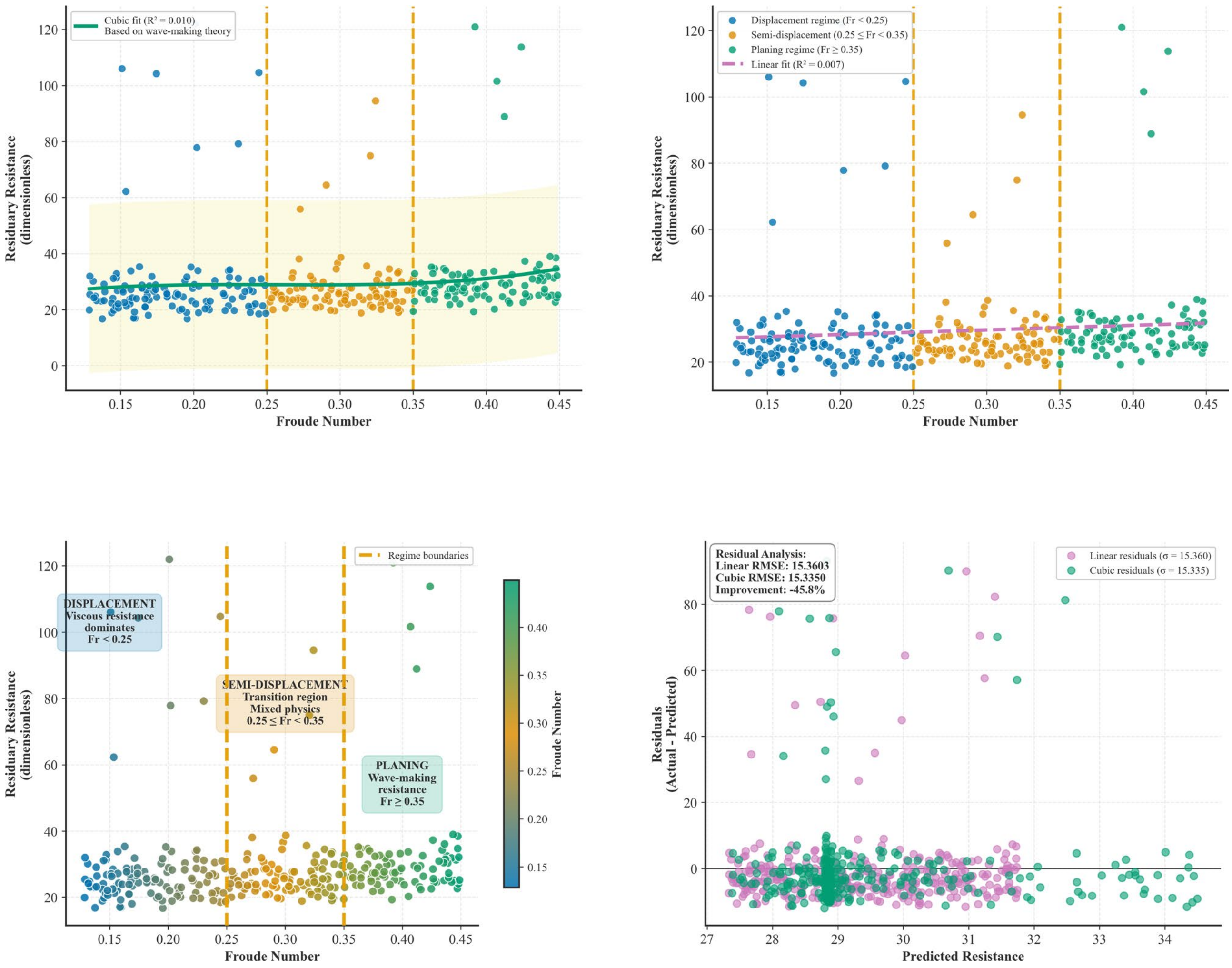
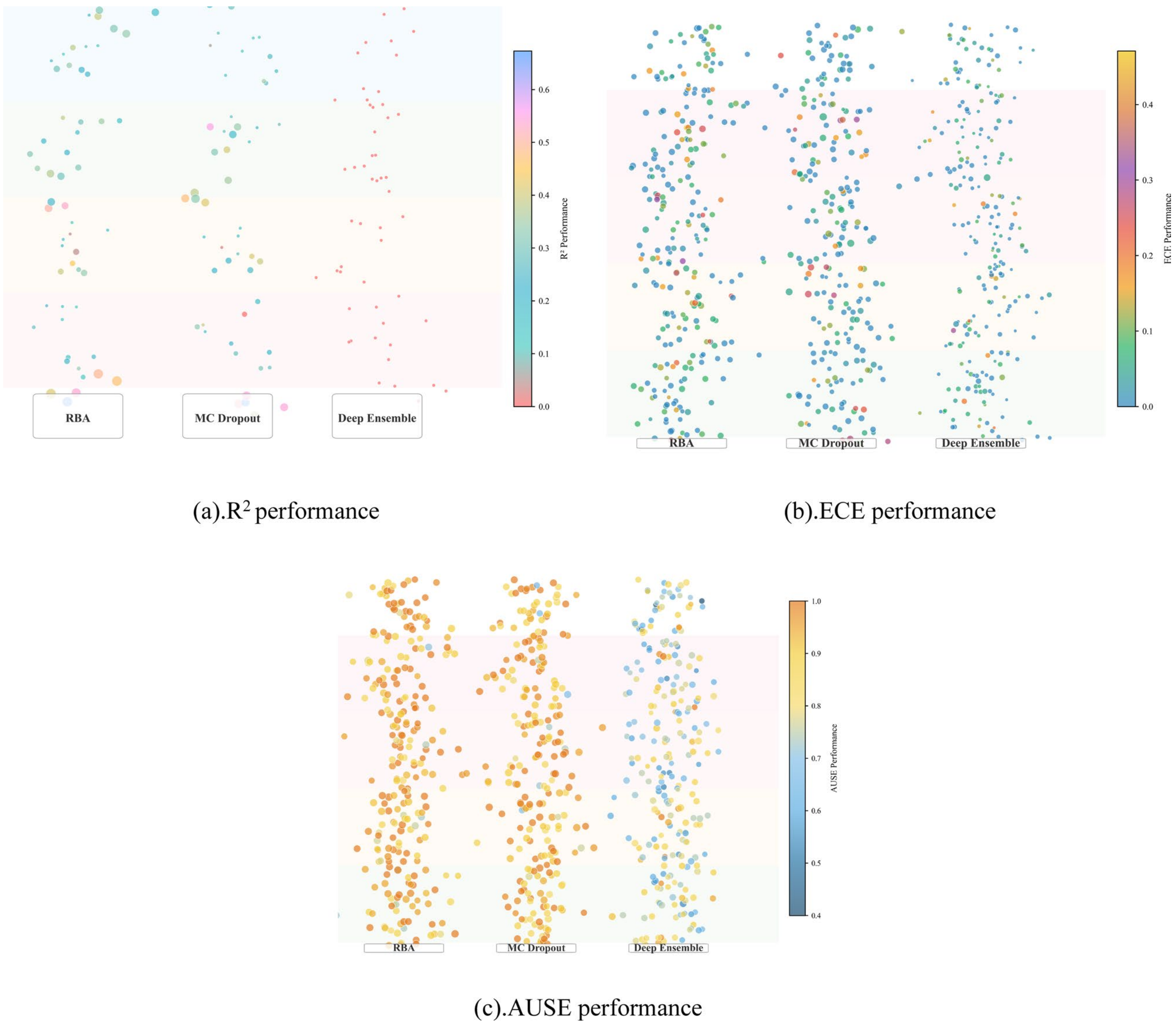


Fig. 12. Froude number complexity analysis



(a).R^2 performance

(b).ECE performance

(c).AUSE performance

Fig. 13. Yacht_Hydrodynamics Dataset in-depth validation of experimental results. (**a**).R2 performance, (**b**). ECE performance, (**c**).AUSE performance

dataset) to stable performance on time series data (Household Power dataset), and then to challenging performance on complex physical systems (Yacht Hydrodynamics dataset). This performance gradient is not simply a matter of method superiority or inferiority, but reveals the fundamental compatibility relationship between RBA's architectural assumptions and different data generating mechanisms. Specifically, when the underlying data structure follows relatively smooth nonlinear mappings and learnable statistical dependencies exist between features, RBA's Bayesian feedforward layers can effectively model parameter uncertainty, the multi-layer residual attention mechanism can capture relevant feature interactions, and the covariance construction module can generate mathematically valid uncertainty representations. However, when the system transitions to multi-physics strongly coupled regime transitions scenarios, the data's generating process fundamentally violates the smoothness assumptions that neural network architectures rely upon, causing all gradient-based optimization methods to face similar predicaments.

More in-depth mechanistic analysis reveals the non-trivial synergistic mechanisms among RBA's three components: ablation experiments show that removal of the Bayesian attention mechanism has the most significant impact on performance, a finding that points to an important architectural insight—in deep probabilistic models, the attention mechanism not only serves the function of feature selection, but more importantly plays a key role as an uncertainty propagation pathway. By directly embedding RBF kernels into attention score computation, RBA achieves a paradigm shift from deterministic similarity measures to probabilistic correlation modeling. This design enables the model to perform principled uncertainty quantification at the feature level, rather than merely post-hoc estimation at the output level. Meanwhile, Bayesian residual connections, through adaptive weighting mechanisms modeled by Beta distributions, achieve inter-layer propagation and accumulation of uncertainty. This design philosophy embodies a profound understanding of unified modeling of information flow and uncertainty flow in deep networks.

From a comparative methodology perspective, RBA's advantages over established baselines such as Monte Carlo Dropout and Deep Ensemble essentially stem from the fundamental difference between its end-to-end uncertainty modeling approach and these methods' approximation-based strategies. MC Dropout approximates weight uncertainty through stochastic regularization, but its uncertainty quality is highly dependent on the specific configuration of dropout rate and network architecture; Deep Ensemble, while capable of providing robust uncertainty estimates, is limited in scalability due to computational overhead and potential diversity-accuracy trade-offs inherent in its ensemble nature. In comparison, RBA achieves a better balance between computational efficiency and theoretical rigor through architectural integration, although this balance still has inherent limitations when facing extremely high complexity tasks.

Particularly worthy of deep consideration is that the modeling challenges of complex physical systems like Yacht Hydrodynamics provide important empirical insights for the entire uncertainty quantification field: when system behavior is governed by multiple, competing physical mechanisms with clear regime boundaries, purely data-driven approaches, regardless of their architectural sophistication, face fundamental expressivity constraints. This observation not only objectively delineates the technical boundaries of current deep learning methods, but also provides clear guidance for future research directions—truly solving such problems requires establishing deeper integration between domain-specific physical priors and general-purpose learning algorithms, which extends far beyond the scope of pure architectural innovation. Therefore, RBA's value lies not only in its practical utility as a working solution, but more importantly in the methodological foundation and empirical benchmarks it provides for understanding and designing next-generation physics-informed probabilistic models.

Conclusion

This study presents the Residual Bayesian Attention (RBA) framework, which integrates Bayesian inference with Transformer architecture through three coupled components: Bayesian feedforward layers, multi-layer residual Bayesian attention, and Bayesian covariance construction. The framework achieves end-to-end probabilistic modeling via variational inference optimization.

Experimental evaluation across six benchmark datasets reveals performance characteristics that depend on data structure and complexity. On structured engineering and time series tasks, RBA demonstrated competitive performance with R^2 ranging from 0.920 to 0.972 and reasonable calibration quality. Ablation experiments confirm that the Bayesian attention mechanism contributes substantively to both accuracy and stability. However, when applied to complex physical systems such as yacht hydrodynamics involving regime transitions across flow regimes, RBA exhibited R^2 values ranging from −27.3% to 3.2%, similar to difficulties encountered by Monte Carlo Dropout and Deep Ensemble methods.

The comparative analysis indicates that RBA provides advantages in prediction interval calibration on medium-scale structured datasets, particularly in maintaining consistency between predicted confidence and actual accuracy. The framework's uncertainty decomposition offers interpretability for decision-support applications, though computational overhead and hyperparameter sensitivity remain practical considerations. These findings suggest that RBA represents a viable solution for uncertainty quantification in scenarios with learnable statistical dependencies and moderate nonlinearity.Future research directions for RBA involve incorporating domain-specific physical priors to improve the model's generalization performance in complex physical systems.

Data availability

The datasets generated and/or analysed during the current study are available in the Hugging Face repository, https://huggingface.co/datasets/guanwencan/Residual-Bayesian-Attention.

Received: 7 August 2025; Accepted: 10 October 2025
Published online: 01 November 2025

References

1. Zhang, J. & Taflanidis, A. A. Multi-objective optimization for design under uncertainty problems through surrogate modeling in augmented input space. *Struct. Multidisciplinary Optim.* **59** (2), 351–372. https://doi.org/10.1007/s00158-018-2069-1 (2019).
2. Zhang, J., Taflanidis, A. A. & Scruggs, J. T. Surrogate modeling of hydrodynamic forces between multiple floating bodies through a hierarchical interaction decomposition. *J. Comput. Phys.* **408**, 109298. https://doi.org/10.1016/j.jcp.2020.109298 (2020).
3. Palar, P. S. et al. *Data-driven Surrogate Modeling Using Deep Learning for Uncertainty Quantification of Random Fields. [Aiaa Scitech 2023 forum]* Jan 23–272023 (AIAA Science and Technology (SciTech) Forum, 2023).
4. Bilbrey, J. A., Firoz, J. S., Lee, M. S. & Choudhury, S. Uncertainty quantification for neural network potential foundation models. *Npj Comput. Mater.* **11** (1). https://doi.org/10.1038/s41524-025-01572-y (2025).
5. Elsharkawy, I., Hooberman, B. & Kahn, Y. Uncertainty quantification from ensemble variance scaling laws in deep neural networks. *Mach. Learning-Science Technol.* **6** (3). https://doi.org/10.1088/2632-2153/adf7fe (2025).
6. Tripathy, R. K. & Bilionis, I. Deep UQ: learning deep neural network surrogate models for high dimensional uncertainty quantification. *J. Comput. Phys.* **375**, 565–588. https://doi.org/10.1016/j.jcp.2018.08.036 (2018).
7. Houichime, T. & El Amrani, Y. Context is all you need: A hybrid Attention-Based method for detecting code design patterns. *Ieee Access.* **13**, 9689–9707. https://doi.org/10.1109/access.2025.3525849 (2025).
8. Huang, X. S., Pérez, F., Ba, J. & Volkovs, M. Jul 13-182020). Improving Transformer Optimization Through Better Initialization.*Proceedings of Machine Learning Research* [International conference on machine learning, vol 119]. International Conference on Machine Learning (ICML), null, ELECTR NETWORK. (2020).
9. Zhang, J., Kailkhura, B. & Han, T. Y. J. Leveraging uncertainty from deep learning for trustworthy material discovery workflows. *ACS Omega.* **6** (19), 12711–12721. https://doi.org/10.1021/acsomega.1c00975 (2021).

10. Liu, B. K. et al. Explainable machine learning for multiscale thermal conductivity modeling in polymer nanocomposites with uncertainty quantification. *Compos. Struct.* **370**, 119292. https://doi.org/10.1016/j.compstruct.2025.119292 (2025).
11. Liu, B. K., Vu-Bac, N., Zhuang, X. Y., Fu, X. L. & Rabczuk, T. Stochastic integrated machine learning based multiscale approach for the prediction of the thermal conductivity in carbon nanotube reinforced polymeric composites. *Compos. Sci. Technol.* 224. https://doi.org/10.1016/j.compscitech.2022.109425 (2022).
12. Liu, B. K., Vu-Bac, N., Zhuang, X. Y., Fu, X. L. & Rabczuk, T. Stochastic full-range multiscale modeling of thermal conductivity of polymeric carbon nanotubes composites: A machine learning approach. *Compos. Struct.* **289**, 115393. https://doi.org/10.1016/j.compstruct.2022.115393 (2022).
13. Rocha, M. B. & Krohling, R. A. VAE-GNA: a variational autoencoder with Gaussian neurons in the latent space and attention mechanisms. *Knowl. Inf. Syst.* **66** (10), 6415–6437. https://doi.org/10.1007/s10115-024-02169-5 (2024).
14. Miao, S. Q., Liu, M. Y. & Li, P. Jul 17-232022). Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism.*Proceedings of Machine Learning Research* [International conference on machine learning, vol 162]. 39th International Conference on Machine Learning (ICML), Baltimore, MD. (2022).
15. Gal, Y. & Ghahramani, Z. Jun 20-222016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.*Proceedings of Machine Learning Research* [International conference on machine learning, vol 48]. 33rd International Conference on Machine Learning, New York, NY. (2016).
16. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Jul 07-092015). Weight Uncertainty in Neural Networks.*Proceedings of Machine Learning Research* [International conference on machine learning, vol 37]. 32nd International Conference on Machine Learning, Lille, FRANCE. (2015).
17. Kingma, D. P., Salimans, T. & Welling, M. Dec 07-122015). Variational Dropout and the Local Reparameterization Trick.*Advances in Neural Information Processing Systems* [Advances in neural information processing systems 28 (nips 2015)]. 29th Annual Conference on Neural Information Processing Systems (NIPS), Montreal, CANADA. (2015).
18. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Dec 04-092017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.*Advances in Neural Information Processing Systems* [Advances in neural information processing systems 30 (nips 2017)]. 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA. (2017).
19. Sensoy, M., Kaplan, L. & Kandemir, M. Dec 02-082018). Evidential Deep Learning to Quantify Classification Uncertainty.*Advances in Neural Information Processing Systems* [Advances in neural information processing systems 31 (nips 2018)]. 32nd Conference on Neural Information Processing Systems (NIPS), Montreal, CANADA. (2018).
20. Gustafsson, F. K., Danelljan, M., Schön, T. B. & Ieee Comp, S. O. C. Jun 14-192020). Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision.*IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* [2020 ieee/cvf conference on computer vision and pattern recognition workshops (cvprw 2020)]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), null, ELECTR NETWORK. (2020).
21. Alemi, A. A., Fischer, I. & Dillon, J. V. Deep Variational Information Bottleneck. *ArXiv*, *abs/1612.00410*. (2017).
22. Tsai, Y. H. H. et al. Multimodal Transformer for Unaligned Multimodal Language Sequences. In A. Korhonen, D. Traum, & L. Màrquez, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* Florence, Italy. (2019), July.
23. Bui, L. M., Huu, T. T., Dinh, D., Nguyen, T. M. & Hoang, T. N. Jul 15-192024). Revisiting Kernel Attention with Correlated Gaussian Process Representation.*Proceedings of Machine Learning Research* [Uncertainty in artificial intelligence]. 40th Conference on Uncertainty in Artificial Intelligence (UAI), Univ Pompeu Fabra, Barcelona, SPAIN. (2024).
24. Louizos, C. & Welling, M. Jun 20-222016). Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors.*Proceedings of Machine Learning Research* [International conference on machine learning, vol 48]. 33rd International Conference on Machine Learning, New York, NY. (2016).
25. Louizos, C. & Welling, M. Aug 06-112017). Multiplicative Normalizing Flows for Variational Bayesian Neural Networks.*Proceedings of Machine Learning Research* [International conference on machine learning, vol 70]. 34th International Conference on Machine Learning, Sydney, AUSTRALIA. (2017).
26. Yang, Y., Huang, J. & Hu, D. X. Sparsify dynamically expandable network via variational dropout. *Appl. Soft Comput.* **160**, Article 111705. https://doi.org/10.1016/j.asoc.2024.111705 (2024).
27. Alvarez, M. A. & Lawrence, N. D. Computationally Efficient Convolved Multiple Output Gaussian Processes. *Journal of Machine Learning Research*, *12*, 1459–1500. ://WOS:000292304000002 (2011).
28. Yu, R., Li, G. Y. & Liu, Y. Apr 09-112018). Tensor Regression Meets Gaussian Processes.*Proceedings of Machine Learning Research* [International conference on artificial intelligence and statistics, vol 84]. 21st International Conference on Artificial Intelligence and Statistics (AISTATS), Lanzarote, SPAIN. (2018).
29. Fashandi, H. Neural module networks: A review. *Neurocomputing* **552**, Article126518. https://doi.org/10.1016/j.neucom.2023.126518 (2023).
30. Finn, C., Abbeel, P. & Levine, S. Aug 06-112017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks.*Proceedings of Machine Learning Research* [International conference on machine learning, vol 70]. 34th International Conference on Machine Learning, Sydney, AUSTRALIA. (2017).
31. Mezzetti, M. & Negri, I. Hierarchical bayesian model to estimate and compare research productivity of Italian academic statisticians. *Scientometrics* **129** (12), 7443–7474. https://doi.org/10.1007/s11192-024-05154-5 (2024).
32. Zhang, F. S. & Gao, C. CONVERGENCE RATES OF VARIATIONAL POSTERIOR DISTRIBUTIONS. *Ann. Stat.* **48** (4), 2180–2207. https://doi.org/10.1214/19-aos1883 (2020).
33. Lee, M. H. Y. Mathematical analysis and performance evaluation of the GELU activation function in deep learning. *J. Math.* **2023** https://doi.org/10.1155/2023/4229924 (2023). Article 4229924.
34. Lipoth, J., Tereda, Y., Papalexiou, S. M. & Spiteri, R. J. A new very simply explicitly invertible approximation for the standard normal cumulative distribution function. *Aims Math.* **7** (7), 11635–11646. https://doi.org/10.3934/math.2022648 (2022).
35. Xu, C. Y. et al. Dec 12–14). Data-Parallel Momentum Diagonal Empirical Fisher (DP-MDEF):Adaptive Gradient Method is Affected by Hessian Approximation and Multi-Class Data. [2022 21st ieee international conference on machine learning and applications, icmla]. 21st IEEE International Conference on Machine Learning and Applications (IEEE ICMLA), Nassau, BAHAMAS. (2022).
36. Bera, A. K. & Koley, M. A history of the delta method and some new results. *Sankhya-Series B-Applied Interdisciplinary Stat.* **85** (2), 272–306. https://doi.org/10.1007/s13571-023-00305-9 (2023).
37. Baek, J. & Han, S. Reinforcement learning using evidence lower bound in variational inference [변분 추론의 증거 하한을 사용한 강화학습]. *J. Inst. Control Rob. Syst.* **28** (11), 981–985. https://doi.org/10.5302/j.Icros.2022.22.0184 (2022).
38. Song, K. et al. Feb 20–27). Robustly Train Normalizing Flows via KL Divergence Regularization.*AAAI Conference on Artificial Intelligence* [Thirty-eighth aaai conference on artificial intelligence, vol 38 no 13]. 38th AAAI Conference on Artificial Intelligence (AAAI) / 36th Conference on Innovative Applications of Artificial Intelligence / 14th Symposium on Educational Advances in Artificial Intelligence, Vancouver, CANADA. (2024).
39. Wen, Z. T. et al. May 12–14). A Lightweight Pipeline Defect Detection Method via Structural Reparameterization Technique and Ghost Convolution.*Data Driven Control and Learning Systems* [2023 ieee 12th data driven control and learning systems conference, ddcls]. IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS), Xiangtan, PEOPLES R CHINA. (2023).
40. Liu, M., Hu, Y. F. & Jin, L. Discrete Data-Driven control of redundant manipulators with adaptive Jacobian matrix. *IEEE Trans. Industr. Electron.* **71** (10), 12685–12695. https://doi.org/10.1109/tie.2023.3347831 (2024).
41. Vaswani, A. et al. Attention Is All You Need. *arxiv*, arXiv:1706.03762. (2017). https://doi.org/10.48550/arXiv.1706.03762

42. Baas, S., Fox, J. P. & Boucherie, R. J. Bayesian covariance structure modeling of interval-censored multi-way nested survival data. *J. Multivar. Anal.* **204**, Article 105359. https://doi.org/10.1016/j.jmva.2024.105359 (2024).
43. Fazio, L. & Buerkner, P. C. Gaussian distributional structural equation models: A framework for modeling latent heteroscedasticity. *Multivar. Behav. Res.* **60** (4), 840–858. https://doi.org/10.1080/00273171.2025.2483252 (2025).
44. Martín, J., Ortiz, W. A. & FOR PROBABILITY METRIC SPACES WITH CONVEX ISOPERIMETRIC PROFILE. SYMMETRIZATION INEQUALITIES *Ann. Academiae Scientiarum Fennicae-Mathematica*, **45**, 877–897. https://doi.org/10.5186/aasfm.2020.4548 (2020).
45. Hao, Z. H., Wang, X. & Ni, X. Y. An eigenvalue decomposition based amplitude extraction method for EIT demodulation system. *Ieee Access.* **12**, 90371–90379. https://doi.org/10.1109/access.2024.3420424 (2024).
46. Thoiyab, N. M. et al. Global stability analysis of neural networks with constant time delay via Frobenius norm. *Math. Probl. Eng.* **2020**, Article4321312. https://doi.org/10.1155/2020/4321312 (2020).
47. Gusev, A., Chervyakov, A., Alexeenko, A. & Nikulchev, E. Particle swarm training of a neural network for the lower upper bound Estimation of the prediction intervals of time series. *Mathematics* **11** (20), 4342. https://doi.org/10.3390/math11204342 (2023).
48. Lv, M. Z. et al. A multi-input and three-output wind speed point-interval prediction system based on constrained many-objective optimization problem. *Inf. Sci.* **720**, 122531. https://doi.org/10.1016/j.ins.2025.122531 (2025).
49. Chen, X., Han, T., Cheng, P. & Da, X. A novel interval prediction method in wind speed based on deep learning and combination prediction. *Sci. Rep.* **15** (1), 23182. https://doi.org/10.1038/s41598-025-03188-0 (2025).
50. Geng, X., Shi, Y., Ou, Y. & Yin, F. Probabilistic forecasting of hand, foot and mouth disease in Mainland China using bayesian additive regression tree model. *Sci. Rep.* **15** (1), 33745. https://doi.org/10.1038/s41598-025-98954-5 (2025).
51. Akram, M. et al. Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by bayesian approaches. *Sci. Rep.* **15** (1), 1342. https://doi.org/10.1038/s41598-024-84478-x (2025).
52. Posocco, N. & Bonnefoy, A. Sep 14-172021). Estimating Expected Calibration Errors.*Lecture Notes in Computer Science* [Artificial neural networks and machine learning - icann 2021, pt iv]. 30th International Conference on Artificial Neural Networks (ICANN), null, ELECTR NETWORK. (2021).
53. Arnold, S., Walz, E. M., Ziegel, J. & Gneiting, T. Decompositions of the mean continuous ranked probability score. *Electron. J. Stat.* **18** (2), 4992–5044. https://doi.org/10.1214/24-ejs2316 (2024).
54. Mansouri, B., Chinipardaz, R., Al-Farttosi, S. A. S. & Mombeni, H. A. A review of nonparametric research on cumulative distribution function Estimation. *J. Indian Soc. Probab. Stat.* **25** (2), 739–760. https://doi.org/10.1007/s41096-024-00201-z (2024).
55. Choudhury, A. & Roy, P. A fairly accurate approximation to the area under normal curve. *Commun. Statistics-Simulation Comput.* **38** (7), 1485–1492. https://doi.org/10.1080/03610910903009344 (2009).

Acknowledgements

We would like to thank Springer Nature Publishing Group for their financial support for this study.We would like to express our special thanks to the editors of Scientific Reports for their efficient handling and patient responses.

Author contributions

Conceptualization: Youliang Chen, Wencan Guan; Methodology: Youliang Chen, Wencan Guan; Software: Youliang Chen (CHG code implementation, Chebyshev confidence module); Validation: Youliang Chen, Wencan Guan; Formal Analysis: Youliang Chen; Investigation: Wencan Guan; Resources: Rafig Azzam; Data Curation: Wencan Guan; Writing – Original Draft: Youliang Chen (Sections: Abstract, Methodology, Case Studies), Wencan Guan (Sections: Introduction, Related Work); Writing – Review & Editing: All authors; Visualization: Youliang Chen; Supervision: Rafig Azzam; Project Administration: Wencan Guan;

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-24093-6.

Correspondence and requests for materials should be addressed to W.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.