Spatially Adaptive Ensemble Learning with Calibrated Predictive Uncertainty

Jeremiah Zhe Liu*[1,3], Sebastian T. Rowland[2], John Paisley[2], Marianthi-Anna Kioumourtzoglou[2], and Brent A. Coull[1]

[1]Harvard University, Boston, MA, USA.
[2]Columbia University, New York, NY, USA.
[3]Google Research, Mountain View, CA, USA.

Abstract

(**jzl:** to update after all edits are done.) Air pollution epidemiology continues to be a critically central area of investigation in the environmental health sciences. Current standard practice in the field is to employ sophisticated spatio-temporal prediction models to estimate exposure to air pollution both at the individual and community levels. A key methodological issue that arises in this context is uncertainty quantification of the resulting exposure estimates. An emerging technique in air pollution exposure assessment is to ensemble several competing exposure prediction models. Ensemble learning is a mainstay in modern data science. Conventional practice assigns to base models a set of deterministic, constant model weights that (1) do not fully account for the individual models' varying accuracy across sub-regions in the feature space and (2) do not provide uncertainty estimates for the ensemble prediction. The first of these shortcomings can yield suboptimal predictions in certain sub-regions of the study, while the second precludes the assessment of prediction reliability and propagation of uncertainty in downstream health effects analyses. We introduce a new Bayesian nonparametric ensemble framework that uses a dependent random measure to adaptively combine models based on their predictive accuracy in the feature space and also nonparametrically models the ensemble's predictive cumulative density function (CDF) so that the model's quantification of the predictive uncertainty is consistent with observed data. We show that the proposed method is asymptotically consistent for uncertainty quantification (i.e. CDF estimation) and can improve upon the predictive performance of an ensemble model with naïve distribution assumptions when the data distribution is complex. We apply the method to data simulated from one- and two-dimensional complex nonlinear regression models, and generate a spatial prediction model and estimate the associated prediction uncertainties for fine particle levels in eastern New England, USA.

Keywords: Ensemble learning; Uncertainty calibration; Gaussian process; Spatiotemporal modeling; Air pollution;

*Work done when first author is a PhD student at Harvard University. This publication was made possible by USEPA grant RD-83587201. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. Funding was also provided by NIEHS grants R01 ES030616, P30 ES000002, and P30 ES009089.

1 Introduction

Ambient air pollution is a well-recognized global environmental threat to human health. In 2015, exposure to $PM_{2.5}$ (particles with aerodynamic diameter ≤ 2.5 μm) was the fifth-ranking mortality risk factor, causing 4.2 million deaths globally [11], and the sixth-ranking factor for disability-adjusted life-years (DALYs), contributing to 103.1 million DALYs [23]. These findings have been confirmed by numerous studies, including three US-wide studies of more than 60 million Medicare beneficiaries that reported significant $PM_{2.5}$ impacts on mortality, even at levels below the current national standards [16, 19, 84].

Despite the long-standing scientific consensus on the harmful effects of $PM_{2.5}$ exposures on the public's health, new questions have been raised about the methodology used to generate this evidence. A *Nature* editorial recently stated this "scientific consensus is now under attack", both in the United Stated and other countries [74]. One major criticism levied against air pollution epidemiological studies is that health effect analyses treat estimated ambient air pollution levels as known [76]. Because it is not feasible—neither cost-wise nor logistically—to ask millions of study participants to constantly wear personal air pollution monitors for long periods of time, studies must rely on estimates of ambient air pollution concentration levels across an area of study.

The current standard practice in the field is to employ sophisticated spatio-temporal prediction models to predict air pollution exposures at residential addresses of study participants, even at locations where the monitoring data are sparse or non-existent [17, 18, 21, 40, 42, 43, 44, 80, 82, 86]. However, due to the difference in data sources and in modeling assumptions, the predictions generated by different air pollution exposure models can differ in space and especially at locations that are far away from the air pollution monitors. As a result, health effect estimates obtained using exposure estimates from a single model can vary, if model uncertainty associated with exposure assignment is not considered. For example, a recent study of the association between long-term $PM_{2.5}$ and coronary artery disease compared different exposure models and found significant effect modification by urban vs. rural residence only for some of the exposure models used, but not all [53]. For the purposes of developing regulatory policy, such model uncertainty is often assessed informally but not quantified rigorously.

In addition to sensitivity of conclusions to choice of exposure model, existing work on methods that properly account for the prediction error when the estimates are used as a covariate in subsequent health effects analyses has shown that failure to adjust for such uncertainty can adversely affect inference on these effects [30, 70, 2]. All of this existing work aims to propagate statistical uncertainty associated with predictions of a single exposure model, which underestimates uncertainty associated with the selection of the exposure model. Further, Alexeeff et al. (2017) showed that use of exposure estimates without accounting for spatially varying prediction error can yield biased health effect estimates. Thus, we seek to quantify both intra- and inter-model uncertainty and how it varies in space, for use in future health effects analyses.

Consequently, to establish rigorous scientific evidence on the health effects of air pollution, it is desirable to produce an ensemble framework for exposure estimation that accomplishes two primary objectives. The first is that it flexibly integrates information from available base exposure models in order to yield the most accurate estimates possible. This objective requires that

the ensemble model should allow ***spatial adaptivity*** in its ensemble weights, which is necessary because some exposure models outperform competitors in some areas but not others [39, 37]. For instance, one model might out-perform another in urban but not rural settings, or near roadways or other microenvironments, or in high but not low concentration settings.

The resulting ensemble must also enable valid quantification of prediction uncertainty. This goal requires that an ensemble model to achieve ***calibrated estimation*** of its predictive uncertainty. In intuitive terms, this means a model's uncertainty estimate reliably reflects the discrepancy between the model prediction and the true observation, such that, e.g., the model's 95% predictive intervals contain new observations 95% of the time [25]. Importantly, to properly quantify predictive uncertainty, the model should recognize different types of uncertainties that arise in the modeling process. In natural science applications, two distinct types of uncertainties exist: *aleatoric uncertainty* and *epistemic uncertainty* [41, 69, 34, 52]. Aleatoric uncertainty arises due to stochastic variability inherent in the data distribution $y|\mathbf{x}$, while the epistemic uncertainty arises due to our lack knowledge or observation of the data-generating mechanism (e.g., predict y at a location $\mathbf{x}$ not well covered by training data). A model's epistemic uncertainty can be reduced by collecting more data, whereas aleatoric uncertainty is irreducible since it is inherent to the data generating mechanism. Consequently, in regions that are well represented by the training data, a model's aleatoric uncertainty should accurately estimate the data-generating distribution by flexibly capturing the stochastic pattern in the data, while in regions unexplored by the training data, the model's epistemic uncertainty should increase to capture the model's lack of confidence in the resulting predictions. The quality of a Bayesian model's uncertainty estimate can be measured rigorously using *proper scoring rules* (e.g., the posterior predictive negative log likelihood $E_{\mu,\sigma}(\frac{(y-\mu(\mathbf{x}))^2}{\sigma(\mathbf{x})})$ under a Gaussian model)[13, 58, 27, 66]. To this end, the model's ability to provide a valid quantification of the aleatoric and epistemic uncertainty is essential for ensuring good performance, especially in out-of-distribution (OOD) regions not well-covered by the training data.

1.1 Contribution to the Model-based Ensemble Literature

In recent years, several ensemble methods have been developed in the environmental literature for air pollution prediction and climate forecasts. Examples include bootstrap aggregation of regression trees or neural networks [48, 33, 81], hierarchical models with sophisticated covariance structure [67, 47], and additive combinations of multiple machine learning predictors [15, 85]. Bayesian variants of the ensemble models also exist, with the well-known examples include those of Smith et al. and Tebaldi et al. ([73, 68]) that place a structured prior on the weights of an additive ensemble, and Gneiting and Weather ([26]), Raftery ([62]), and Murray ([59]) that perform Bayesian model averaging (BMA) on the base models' predictive distributions. However, these existing approaches either do not allow spatial adaptivity in the ensemble weights or assume that the true data generating model lies strictly within the convex combination of the base model predictions (e.g., the $\mathscr{M}$-closed assumption made by BMA)[10].

In this work, we present a principled Bayesian approach to ensemble learning that achieves both a ***spatially adaptive ensemble*** and ***calibrated predictive uncertainty***. That is, it adaptively combines predictions from multiple spatio-temporal processes, addresses model biases in the

predictive distribution assumptions using Bayesian nonparametric machinery, and achieves a complete quantification of both the aleatoric and epistemic uncertainty. Specifically, we model the ensemble weights as a spatially dependent random measure, and then couple the ensemble model's prediction function with a residual process that mitigates the systematic bias shared by the base models. We also model the ensemble's distribution function semi-parametrically as a random function, carefully specifying its priors so that the posterior predictive distribution flexibly captures the data's empirical distribution while maintaining statistical efficiency via shrinkage mechanisms. We term our method *Adaptive Bayesian Nonparametric Ensemble* (BNE). From the perspective of statistical modeling, BNE can be understood as a $\mathscr{M}$-open generalization of BMA, where instead of restricting the model's distribution function to be a convex combination of the base model predictions, BNE uses its Bayesian nonparametric machinery to generate a large model space for the distribution function that is *a priori* centered at that of the original ensemble. As a result, even when the base models are misspecified, BNE is able to achieve improved predictive accuracy and more calibrated predictive intervals.

1.2 The Data: $PM_{2.5}$ in Eastern New England, USA

Eastern New England, USA is an area in which many different air pollution cohort studies are on-going, including cohorts of children (Project Viva), middle aged populations (Framingham Heart Study), and the elderly (the Normative Aging Study), among others [36, 22, 20]. Figure 1 shows the average of daily $PM_{2.5}$ predictions in eastern New England during 2011 from three published $PM_{2.5}$ prediction models [42, 80, 18]. Although all three models reported roughly similar predictive accuracy in the broader study areas for which they were developed, they yielded different predictions in eastern New England, the focus of our case study. This highlights the importance of adaptive weights in any ensemble model based on these predictions. We provide some detail on the data and algorithms that yielded each set of predictions here.

The first model (top left of Figure 1) was trained on monitoring data in the Northeastern United States [42]. It used a hybrid satellite-based model incorporating daily satellite remote sensing data, monitored concentrations, meteorological variables and land use regression variables. This model was validated using cross-validation (CV) of monitoring data, achieving an out-of-sample R^2 of 0.88 for days with all of the predictor information and 0.87 for days without. The second model (top middle of Figure 1) predicts average $PM_{2.5}$ during 2011 from a convolutional neural net trained on monitoring data from 2000–2012 over the continental U.S. [18]. The model incorporates remote sensing satellite data, simulated values from a chemical transport model, and land-use and meteorological terms. Ten-fold cross-validation yielded an out-of-sample of R^2 of 0.84 on left-out monitoring data across the country. The third model (top right of Figure 1) shows average predicted $PM_{2.5}$ from an approach that used geographically weighted regression (GWR) to correct bias arising from a model based on remote sensing satellite data on aerosol optical depth (AOD) and a chemical transport model [80]. The authors report this approach yielded out-of-sample R^2 of 0.82 over North America during the period of 2004-2008. All three models provided predictions at a 1×1 km grid resolution.

Taken together, these models incorporate different sources and forms of information on ambient $PM_{2.5}$ levels and are applied at different spatial scales. Going forward, we refer to the

models by first author: the Kloog model (Figure 1, Top Left), the Di model (Figure 1, Top Middle), and the van Donkelaar model (Figure 1, Top Right).

We apply BNE to integrate the Kloog, Di, and van Donkelaar models for $PM_{2.5}$ concentrations for 2011 in eastern New England, USA. In this application, BNE generated $PM_{2.5}$ predictions that outperform those of existing approaches in terms of out-of-sample predictive accuracy. It also produced *calibrated* estimates of predictive uncertainty, producing predictive intervals that have much lower negative log likelihood (NLL) and expected coverage error (ECE) than the competing methods. Furthermore, we show that by leveraging the posterior estimate of BNE's different model components, we can decompose the model's overall predictive uncertainty into different sub-components (i.e., those due to ensemble weight estimation and those due to prediction) to understand the factors driving overall model uncertainty.

2 Model

Adaptive Bayesian Nonparametric Ensemble (BNE) generalizes a classic ensemble model by augmenting the model's parametric components with Bayesian nonparametric machinery. Given observations $\{y_i, \mathbf{x}_i\}_{i=1}^n$ where $y \sim F^*(y|\mathbf{x})$ and a set of deterministic model predictions $\{f_k\}_{k=1}^K$, a classic ensemble model assumes the form:

$$Y = \sum_{k=1}^{K} f_k(\mathbf{x})\omega_k + \varepsilon, \tag{1}$$

where $\omega = \{\omega_k\}_{k=1}^K$ are the ensemble weights assigned to each base model, and ε is a zero-mean random variable describing the distribution of the outcome noise, which is commonly assumed to be Gaussian $\varepsilon \sim N(0, \sigma^2)$.

The prediction of this classic Bayesian ensemble model is assessed using its posterior predictive distribution. Denoting $\Phi_\sigma(y|\mathbf{x}, \mu)$ the cumulative distribution function (CDF) of a Gaussian distribution, the posterior predictive distribution of a classic ensemble model is obtained by integrating $\Phi_\sigma(y|\mathbf{x}, \mu)$ over the posterior distribution of the model parameter μ:

$$F_{\mu,\sigma}(y|\mathbf{x}) = \int_{\mathbb{R}} \Phi_\sigma(y|\mathbf{x}, \mu)\pi_n(\mu, \sigma)d\mu d\sigma \quad \text{where} \quad \mu = \sum_{k=1}^{K} f_k(\mathbf{x})\omega_k, \tag{2}$$

where $\pi_n(\mu, \sigma)$ denotes the posterior distribution (μ, σ) given $\{y_i, \mathbf{x}_i\}_{i=1}^n$, and we have denoted the predictive distribution function as $F_{\mu,\sigma}$ to stress its dependency on both the specification of the mean function $\mu = \sum_k f_k(\mathbf{x})\omega_k$ and of the variance σ. Then, the model's point prediction is expressed using the posterior mean $E(y|\mathbf{x}) = \int y dF_{\mu,\sigma}(y|\mathbf{x})$, and the model's predictive uncertainty is expressed using the $(1-q)\%$ level credible interval $[F^{-1}_{\mu,\sigma}(\frac{q}{2}|\mathbf{x}), F^{-1}_{\mu,\sigma}(1-\frac{q}{2}|\mathbf{x})]$. Clearly, the quality of the predictive distribution $F_{\mu,\sigma}(y|\mathbf{x})$ depends both on the specification of mean function μ and the appropriateness of the distribution assumption in Φ_σ. As a result, model biases in μ and Φ_σ negatively impact not only an ensemble model's predictive accuracy, but also its reliability in the quantifying the model's predictive uncertainty for the outcome y.

To achieve spatial adaptivity and mitigate model biases, BNE augments both the mean and the predictive distribution function of a classic ensemble model using Bayesian nonparametric

machinery. For the mean function μ, BNE replaces the constant weights $\{\omega_k\}_{k=1}^K$ with random measures $\{\omega_k(\mathbf{x})\}_{k=1}^K$ so it combines the base predictors differently depending on the location of the input feature $\mathbf{x}$ in the feature space, and then adds to μ a flexible residual process $\delta(\mathbf{x})$ to mitigate the systematic prediction bias shared among all the base predictors f_k's. As a result, the mean function becomes:

$$\mu = \sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x}) + \delta(\mathbf{x}).$$

Then, to mitigate model biases in quantifying predictive uncertainty, BNE models its predictive distribution function as a nonparametric random function G centered around the original $F_{\mu,\sigma}$. So the final form of the BNE is:

$$y|\mathbf{x} \sim G\big[F_{\mu,\sigma}(y|\mathbf{x})\big] \quad \text{where} \quad \mu = \sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x}) + \delta(\mathbf{x}). \tag{3}$$

In particular, the random function G can be considered as a flexible "calibration function" that transforms the original parametric predictive CDF $F_{\mu,\sigma}(y|\mathbf{x})$ toward the true data-generating CDF $F^*(y|\mathbf{x})$ underlying the observed data, thereby relaxing the constant-variance Gaussian assumption encoded in Φ_σ. The model parameters ω (ensemble weights), δ (residual process) and G (calibration function) are random functions following suitable Bayesian nonparametric priors. In the following, Section 2.1 introduces the prior specifications for the main model parameters (ω, δ, G). Section 2.2 discusses the choices for the kernel functions and model hyperparameters. Finally, Section 3 derives the expressions for the BNE's predictive mean and variances in terms of model parameters, thereby making it explicit the roles that each model parameter plays in model prediction and uncertainty quantification.

2.1 Model Specification

2.1.1 Mean Function: Spatially Adaptive Ensemble

Recall that to achieve spatial adaptivity, BNE's mean function $\mu = \sum_{k=1}^K f_k(\mathbf{x})\omega_k(\mathbf{x}) + \delta(\mathbf{x})$ consists of the spatially adaptive ensemble weight ω, and a flexible residual process δ to mitigate the spatially varying systematic bias. The priors for ω and δ are specified as:

$$\omega \sim \text{Tailfree}(W, k_\omega), \quad \delta \sim GP(0, k_\delta), \tag{4}$$

respectively. Specifically, $\omega_k(\mathbf{x})$ is a collection of random measure that controls the contribution of each individual base model f_k to the overall ensemble. We model ω as the multivariate logistic transformation of K independent Gaussian processes corresponding to each f_k:

$$\omega_k(\mathbf{x}) = \frac{exp\big(w_k(\mathbf{x})\big)}{\sum_{k'=1}^K exp\big(w_{k'}(\mathbf{x})\big)}, \qquad \{w_k\}_{k=1}^K \overset{iid}{\sim} GP(\mathbf{0}, k_\omega). \tag{5}$$

Denoting the collection of Gaussian processes $\{w_k\}_{k=1}^K$ as W, we have specified a dependent tail-free process (DTFP) prior [35] for the ensemble weights, which we denote as $\omega \sim \text{Tailfree}(W, k_\omega)$.

We model δ nonparametrically using a Gaussian process (GP) with zero mean function $\mathbf{0}(x) = 0$ and kernel function $k_\delta(\mathbf{x}, \mathbf{x}')$. As a result, in densely-sampled regions that are well captured by the training data, $\delta(\mathbf{x})$ will confidently (i.e., with low posterior variance) mitigate the prediction bias between the observation y and the prediction function $\sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x})$. However, in sparsely-sampled regions, the posterior mean of $\delta(\mathbf{x})$ will fall back to $\mathbf{0}(x) = 0$ so as to leave the predictions of the original ensemble intact (since these expert-built base models presumably have been specially designed for the problem being considered), and the posterior uncertainty of $\delta(\mathbf{x})$ will increase to reflect the model's increased uncertainty at location $\mathbf{x}$ that is far away from the observations.

2.1.2 Distribution Function: Heteroskedastic Exponentially-modified Gaussian Process (HEX-GP)

To flexibly capture the complex, spatially-heterogeneous patterns in the air pollution distribution, BNE models its predictive distribution $F(y|\mathbf{x})$ as a flexible transformation of the original predictive distribution function $F_{\mu,\sigma}$:

$$F(y|\mathbf{x}) = G\big[F_{\mu,\sigma}(y|\mathbf{x})\big], \tag{6}$$

where G is a spatially-adaptive "calibration function" following a suitable Bayesian nonparametric prior.

To design a proper G model, we notice that in air pollution modeling, the data observations (e.g., measurements from air pollution monitors) are distributed sparsely and unevenly across the geographical regions, yet the total sample size can be large due to the number of total monitors and, potentially, repeated measurements. As a result, the preference is put on parsimonious models that offers both computationally tractability and sufficient expressiveness to capture the spatially-varying patterns of dispersion and skewness.

To this end, we present a tractable extension to the classic Gaussian process which we term *Heteroskedastic Exponentially-modified Gaussian process* (**HEX-GP**). **HEX-GP** augments a Gaussian process model with flexible Bayesian nonparametric estimators for its higher moments, and deploys smooth shrinkage on these moment estimators to guard against overfitting. Specifically, given a vanilla Gaussian process model

$$y|\mathbf{x} \sim N(\mu(\mathbf{x}),\, \sigma^2), \quad \mu(\mathbf{x}) \sim GP(\mathbf{0}, k_\mu),$$

HEX-GP models the spatially-heterogeneous variance in $y|\mathbf{x}$ by introducing a non-negative scaling factor $s(\mathbf{x})$ to its variance component, such that:

$$y|\mathbf{x} \sim N(\mu(\mathbf{x}),\, \sigma^2 * s^2(\mathbf{x}));$$

where $s(\mathbf{x})$ follows a Bayesian non-parametric prior that is *a priori* centered around 1. Then, to model the potential patterns of skewness in data distribution, we introduce an *exponential-modification* $m(\mathbf{x})$ to the mean function:

$$y|\mathbf{x} \sim N(\mu(\mathbf{x}) + m(s),\, \sigma^2 * s^2(\mathbf{x})); \quad m(\mathbf{x}) \sim Exp(\lambda(\mathbf{x})). \tag{7}$$

where $\lambda(\mathbf{x})$ is the scale parameter of the exponential distribution. Notice that if holding (μ, σ, s) constant, (7) leads to the well-known *Exponentially-modified Gaussian* (EMG) distribution for skewed continuous data on the real support $\mathbb{R}$ [29]. The EMG distribution is able to exhibit a Pareto-like (fat) right tail, and reverts back to a Gaussian shape as $\lambda \to 0$ [65].

Under (7), we specify the following log priors for (s, λ):

$$log\,\lambda \sim GP(\mu_\lambda, \sigma_\lambda^2 * k_\lambda); \quad log\,s \sim GP(\mathbf{0}, \sigma_s^2 * k_s); \tag{8}$$

where μ_λ is a large negative constant that centers $\lambda(\mathbf{x})$ *a priori* to Δ, a small value close to 0. Here $(\sigma_\lambda^2, \sigma_s^2)$ are pre-specified hyper-priors that control the strength of model shrinkage of the nonparametric estimators $(m(\mathbf{x}), s(\mathbf{x}))$ towards their prior centers Δ and 1.

Coming back to (6), the **HEX-GP** induces a calibration function G that takes an tractable form:

$$G\big[F_\mu(y|\mathbf{x})\big] = F_\mu(y|\mathbf{x}) - C(\mathbf{x}) * e^{-\frac{y-\mu(\mathbf{x})}{\lambda(\mathbf{x})}} * F_\mu\Big(y - \frac{\sigma^2 s^2(\mathbf{x})}{\lambda(\mathbf{x})}\Big|\mathbf{x}\Big)$$

where $C(\mathbf{x}) = exp\big(\frac{\sigma^2 s^2(\mathbf{x})}{2\lambda^2(\mathbf{x})}\big)$ is a positive constant. As shown, G applies a sophisticated additive modification to the original distribution function F_μ, shifting the probability mass toward the positive direction by reducing the value of F_μ when y is small. The degree of modification is modulated by the magnitude of λ, and disappears as $\lambda \to 0$.

2.2 Kernel Functions and Hyperparameters

Throughout this work, we model k_ω, k_δ and (k_s, k_λ) using the Matérn $\frac{3}{2}$ kernel family with a length-scale parameter $l \geq 0$:

$$k(\mathbf{z}, \mathbf{z}'|l) = \Big(1 + \frac{\sqrt{3}||\mathbf{z}-\mathbf{z}'||_2}{l}\Big) * exp\Big(-\frac{||\mathbf{z}-\mathbf{z}'||_2}{l}\Big). \tag{9}$$

We denote the length-scale parameters for k_ω, k_δ and k_G as l_ω, l_δ, and l_G, respectively.

Matérn $\frac{3}{2}$ kernels are well suited for quantifying the predictive uncertainty of a given model, since they produce predictive variances that are explicitly characterized by the distance of a given point from the training data. Therefore, for a model with a Matérn $\frac{3}{2}$ kernel, its predictive variances increases as the prediction location moves further away from the training data. Furthermore, a Matérn $\frac{3}{2}$ kernel comes with a theoretical guarantee in estimating distribution parameters. That is, the sample space of a Matérn $\frac{3}{2}$ process corresponds to the space of Hölder continous functions that are at least once differentiable, allowing (δ, s, λ) to flexibly capture the spatially adaptive biases and overdispersions in the data [77, 79].

The hyperparameters for a BNE are σ^2 (the noise variance) the kernel hyperparameters $\{l_\omega, l_\delta, l_s, l_\lambda\}$, and the distribution-parameter regularizers $(\mu_\lambda, \sigma_\lambda, \sigma_s)$. In practice, to guarantee a proper rate of posterior convergence, it is important that the model hyperparameters are estimated adaptively [78]. Consistent with the existing GP approaches [72], we place the inverse-Gamma priors on the l's and the Half Normal priors on σ, i.e.:

$$l_\theta \sim invGamma(\alpha_\theta, \beta_\theta) \quad \text{for } \theta \in \{\omega, \delta, s, \lambda\},$$

where α, β where chosen so the prior probability for l falling into a desired range $[l_{lower}, l_{upper}]$ is high. In this work, we set $l_{lower} = 2$ and $l_{upper} = 10$ such that $P_{invGamma}(l \in [2,10] | \alpha, \beta) = 0.98$ for all length-scale parameters. For the variance parameter σ, we use the weakly informative half-Gaussian prior:

$$\sigma \sim HalfNormal(0,5).$$

Finally, to regularize the variance and skewness parameters (s, λ), we set $\mu_\lambda = -3$ and

$$\sigma_s \sim HalfNormal(0,1), \quad \sigma_\lambda \sim HalfNormal(0,1),$$

thereby encourage shrinkage on (s, λ) by putting nontrivial prior mass around zero.

3 Prediction and Uncertainty Quantification

The model parameters for BNE are ω (the ensemble weights), δ (the residual process) and $G = (s, \lambda)$ (the calibration function). Denoting $\bar{\beta}$ the posterior mean of a BNE model parameter $\beta \in \{\omega, \delta, s, \lambda\}$, the predictive mean of the BNE is:

$$\begin{aligned} E(y|\mathbf{x}) &= \sum_{k=1}^{K} f_k(\mathbf{x})\bar{\omega}_k(\mathbf{x}) + \bar{\delta}(\mathbf{x}) + \int_{y \in \mathscr{Y}} \Big[F_\mu(y|\mathbf{x}) - \bar{G}\big[F_\mu(y|\mathbf{x})\big]\Big] dy \\ &= \sum_{k=1}^{K} f_k(\mathbf{x})\bar{\omega}_k(\mathbf{x}) + \bar{\delta}(\mathbf{x}) + \bar{\lambda}(\mathbf{x}) \end{aligned} \tag{10}$$

(see Section A for a derivation of this fact). This results shows that the predictive mean for a BNE is composed of three components: (1) the predictive mean of the original adaptive ensemble $\sum_{k=1}^{K} f_k(\mathbf{x})\bar{\omega}(f_k, \mathbf{x})$; (2) the term $\bar{\delta}$ representing BNE's "direct" correction to the prediction function obtained through estimation of the residual process δ, and (3) the term $\int \big[F_\mu(y|\mathbf{x}) - \bar{G}[F_\mu(y|\mathbf{x})]\big] dy = \bar{\lambda}(\mathbf{x})$ representing a BNE's "indirect" correction to the prediction function obtained upon relaxation of the Gaussian assumption.

To understand the behavior of BNE's model uncertainty, we decompose $Var(y|\mathbf{x})$ using law of total variance:

$$\begin{aligned} Var(y|\mathbf{x}) &= \underbrace{E\big[Var(y|\mathbf{x}, (\omega, \delta, G))\big]}_{aleatoric\ \ uncertainty} + \underbrace{Var\big[E(y|\mathbf{x}, (\omega, \delta, G))\big]}_{epistemic\ \ uncertainty} \\ &= E\Big[\sigma^2 s^2(\mathbf{x}) + \lambda^2(\mathbf{x})\Big] + Var\Big[\sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x}) + \delta(\mathbf{x}) + \lambda(\mathbf{x})\Big]. \end{aligned} \tag{11}$$

As shown, given a location $\mathbf{x}$, the first component $E\Big[\sigma^2 s^2(\mathbf{x}) + \lambda^2(\mathbf{x})\Big]$ describes the model's posterior estimate of the *aleatoric uncertainty* (i.e., the observed stochasticity in the data distribution), whereas the second component $Var\big[\sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x}) + \delta(\mathbf{x}) + \lambda(\mathbf{x})\big]$ is the model's posterior variance reflecting the *epistemic uncertainty*, i.e., the uncertainty due to the lack of similar examples in the training data [14, 69, 34]. As a result, in the regions well represented by

the training data, the model flexibly captures the heterogeneous patterns in the data distribution while maintaining a low level of epistemic uncertainty. On the other hand, in regions outside the training domain, the aleatoric uncertainty reverts back to σ^2 as the adaptive components reverting to their prior $(s,\lambda) \to (1,\Delta)$, while the epistemic uncertainty increases as δ reverts back to their Gaussian process priors with high variance.

From the perspective of uncertainty quantification, each component of (11) plays a crucial role: without the aleatoric component, the model loose the ability to capture the spatially adaptive patterns in the data due to variance (via $s^2(\mathbf{x})$) or heavy-tailness (via $\lambda^2(\mathbf{x})$). In this case, the model confidence interval will only be proportional to the model's epistemic uncertainty. This can be problematic in well-observed regions with above-average level of data noise, where the confidence interval would fail to achieve its nominal coverage due to low-degree of epistemic uncertainty. On the other hand, without the epistemic component (esp. the posterior variance in $\delta(\mathbf{x})$), the model loses the ability to express heightened uncertainty in regions under-represented in the training data, which may lead to severely over-confident predictive intervals that is too narrow to cover the true observations.

4 Model Fitting and Inference

For a Bayesian model that specifies a model space of predictive distribution $\mathscr{M} = \{F(y|\mathbf{x})\}$, model estimation finds the projection of the empirical distribution $\mathbb{F}(y|\mathbf{x})$ of the data in the model space $\mathscr{M}$. To this end, an BNE specifies a large model space that is centered around a naive ensemble model $Y = \sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x}) + \varepsilon$. Specifically, as shown in Figure 2, an BNE starts from the naive model space assuming no prediction bias and a Gaussian outcome:

$$\mathscr{M}_{\omega} = \left\{ F(y|\mathbf{x}) \middle| F = F_{\mu}(y|\mathbf{x}), \mu = \sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x}) \right\}.$$

Then, by adding δ, an BNE expands $\mathscr{M}_{\omega}$ to a model space $\mathscr{M}_{\omega,\delta}$ using the bias-corrected (i.e. nonparametric) prediction function μ, still under the Gaussian assumption for the outcome:

$$\mathscr{M}_{\omega,\delta} = \left\{ F(y|\mathbf{x}) \middle| F = F_{\mu}(y|\mathbf{x}), \mu = \sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x}) + \delta(\mathbf{x}) \right\}.$$

Finally, using G, an BNE expands to the full model space $\mathscr{M}_{\omega,\delta,G}$, which represents a nonparametric prediction function based on minimal distribution assumptions:

$$\mathscr{M}_{\omega,\delta,G} = \left\{ F(y|\mathbf{x}) \middle| F = G\left[F_{\mu}(y|\mathbf{x})\right], \mu = \sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x}) + \delta(\mathbf{x}) \right\}.$$

It is easy to see that the three model spaces are nested, i.e. $\mathscr{M}_{\omega} \subset \mathscr{M}_{\omega,\delta} \subset \mathscr{M}_{\omega,\delta,G}$.

Notice that for nonoparametric estimation, naively projecting $\mathbb{F}(y|\mathbf{x})$ to the full model space $\mathscr{M}_{\omega,\delta,G}$ (i.e. fitting the full BNE model directly) can lead to non-identifiability of the model parameters. In particular, since ω and δ are non-identifiable in the mean component μ, estimation of the full BNE model directly can result in the undesirable situation where the flexible residual

process δ overfits the training data and "squeezes" the estimates for the ensemble weight ω to a suboptimal value, which not only negatively impacts the model's prediction performance, and also makes it difficult to distinguish between the discrepancy of the base models and the effects of the nonparametric calibration parameters [3, 6, 75].

Consequently, to ensure the identifiability, in this work we propose a step-wise algorithm that performs model estimation by finding projections in a sequentially expanding model space $\mathscr{M}_\omega \subset \mathscr{M}_{\omega,\delta} \subset \mathscr{M}_{\omega,\delta,G}$, thereby placing implicit regularization on model parameters in each step. We show that the proposed algorithm is computationally convenient in that it decomposes the estimation of the full BNE into three classic sub-problems (Section 4.1).

4.1 The Algorithm

Briefly, the algorithm first initializes δ and G to their default values $\delta(\mathbf{x}) = 0$ (the zero function) and $G(\mathbf{z}) = \mathbf{I}(\mathbf{z})$ (the identity function, this is done by setting $(\lambda, s) = (0, 1)$), then updates the model parameters sequentially. That is, it updates ω holding $\delta = \mathbf{0}, G = \mathbf{I}$, then update $\delta|\omega$ holding $G = \mathbf{I}$, and so on. The approach first finds the ensemble weights ω that best explain the data with a model assuming no prediction bias (i.e. $\delta = \mathbf{0}$) and under the Gaussian assumption for the outcome (i.e. $G = \mathbf{I}$), then estimates $\delta(\mathbf{x})$ to correct for residual bias under the Gaussian assumption, and finally estimates G to relax the Gaussian assumption for the outcome. The algorithm concludes after estimating G. Specifically:

Step 1: **Adaptive Ensemble**: **Update** ω

This step fits the naive ensemble model $Y = \sum_{k=1}^{K} f_k(\mathbf{x})\omega_k(\mathbf{x}) + \varepsilon$, as specified by $\mathscr{M}_\omega$. The model likelihood is:

$$f(\omega|\{y_i, \mathbf{x}_i\}) \propto exp\bigg(-\frac{1}{2\sigma^2}\Big(y_i - \sum_{i=1}^{k} f_k(\mathbf{x}_i)\omega(\hat{f}_k, \mathbf{x})\Big)^2\bigg) \prod_{k=1}^{K} f(w_k),$$

where $f(w_k)$ are the Gaussian process likelihoods for the DTFP prior of ω in (5).

Since $\{f_k(\mathbf{x}_i)\}_{k=1}^{K}$ are deterministic, this step can be understood as a linear regression but with the regression coefficients ω adaptive to the input $\mathbf{x}_i$ [61].

Step 2: **Bias Correction**: **Update** $\delta|\omega$

This step expands the model space from $\mathscr{M}_\omega$ to $\mathscr{M}_{\omega,\delta}$ by fitting the bias-corrected ensemble model conditional on ω:

$$Y = \sum_{k=1}^{K} f_k(\mathbf{x})\omega(f_k, \mathbf{x}) + \delta(\mathbf{x}) + \varepsilon. \tag{12}$$

This step can be understood as running a Gaussian process regression. This is because conditional on ω, (12) is a Gaussian process model with mean function $\mu_\omega = \sum_{k=1}^{K} f_k(\mathbf{x})\omega(f_k, \mathbf{x})$ and

kernel function k_δ. The posterior distribution for $\delta|\omega$ is a Gaussian process:

$$\begin{aligned}
&\delta|\omega \sim GP\Big(\mu_{\delta|\omega}, \mathbf{K}_{\delta|\omega}\Big), \qquad \text{where} \\
&\mu_{\delta|\omega} = \mu_\omega + \mathbf{K}_\delta(\mathbf{K}_\delta + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mu_\omega) \qquad \text{and,} \\
&\mathbf{K}_{\delta|\omega} = \mathbf{K}_\delta - \mathbf{K}_\delta(\mathbf{K}_\delta + \sigma^2\mathbf{I})^{-1}\mathbf{K}_\delta,
\end{aligned}$$

where $\mathbf{K}_\delta$ is the kernel matrix constructed using kernel function k_δ.

Step 3: **Distribution Calibration**: **Update** $G, \delta|\omega$

This step expands $\mathscr{M}_{\omega,\delta}$ to the full model space $\mathscr{M}_{\omega,\delta,G}$ by estimating the calibrated predictive distribution function $F(y|\mathbf{x}) = G[F_\mu(y|\mathbf{x})]$, where $F_\mu(y|\mathbf{x})$ is the predictive distribution function generated by the bias-corrected ensemble model under Gaussian distribution assumption from Step 2.

To estimate $G = (s, \lambda)$ and update δ, we may simply initialize from $\mathscr{M}_{\omega,\delta}$, and run a standard inference algorithm with respect to the hierarchical model:

$$\begin{aligned}
y|\mathbf{x} &\sim N(\mu_\omega(\mathbf{x}) + \delta(\mathbf{x}) + m(\mathbf{x}), \sigma^2 * s^2(\mathbf{x})) \\
\delta &\sim GP(0, k_\delta), \quad log\, s \sim GP(0, \sigma_s^2 * k_s) \\
m &\sim Exp(\lambda), \quad log\, \lambda \sim GP(\mu_\lambda, \sigma_\lambda^2 * k_\lambda),
\end{aligned} \tag{13}$$

where (μ_ω, σ^2) follows the posterior distribution from $\mathscr{M}_\omega$, and $(\sigma_s^2, \sigma_\lambda^2)$ are the pre-specified hyper-parameters controlling the strength of model skrinkage onto (λ, s). Notice that under BNE, the mean-parameter convolution $\delta(\mathbf{x}) + m(\mathbf{x})$ does not lead non-identifiability issues. This is because the λ parameter is controlled by the skrinkage prior (8) and can be uniquely identified in the presence of skewness and heavy-tail in the data distribution.

A comment regarding computation is in order. As shown, the proposed algorithm decomposes the full estimation problem into three sub-problems, where the model progressively imposes layers of Bayesian nonparametric augmentations to address potential misspecification while maintaining identifiability. The posterior distributions for each of these sub-problems either have closed-form, or are smooth and differentiable functions of model parameters that can be sampled efficiently using standard gradient-based MCMC methods such as the No-U-Turn Sampler (NUTS) [32]. However, for larger data sets, the issue of computational scalability need to be addressed. For example, the sample size of a spatio-temporal dataset is typically many times larger than that of a spatial dataset, which presents significant computational challenges to a GP due to its computational complexity of $O(n^3)$. To this end, in Appendix B we supply a practical random-feature method that approximates the GP posterior with $O(n)$ computational complexity [63]. This approach, combined with modern minibatch-based stochastic gradient MCMC algorithms, allow the method to be easily scalable to millions of observations [4].

5 Simulation Studies

(**jzl:** yet to finish updating this section, please skip.) (**jzl:** TODO: Update Figures, add baseline models.)

We investigate the performance of the proposed method on a collection of simulated time-series and spatial regression datasets, and consider rigorous metrics to comprehensively evaluate the model out-of-sample behavior in prediction and uncertainty calibration. In particular, we measure the model's predictive accuracy using the mean squared error (MSE), and we measure the model quality in uncertainty calibration with two metrics: the negative log likelihood (NLL) of the model's posterior predictive distribution, and also the expected coverage error (ECE), which measures the empirical coverage of the model's predictive intervals.

Specifically, to compute NLL, we notice that all models in the experiment can be written in form that adopts an Gaussian outcome. As a result, given p the joint posterior distribution of (μ, σ), the posterior predictive log likelihood is:

$$NLL = \int_{\mu,\sigma} \Big[\underbrace{\frac{(y-\mu(x))^2}{\sigma^2}}_{calibration} + \underbrace{log\sigma}_{sharpness} \Big] dp(\mu,\sigma). \tag{14}$$

As shown, NLL simultaneously measures the model prediction's quality in *calibration* (i.e., if the model's predictive uncertainty correspond well with the prediction error, as measured by $(y-\mu(x))^2/\sigma^2$) and in *sharpness* (i.e., if the model's predicted intervals are reasonably tight, as measured by $log\,\sigma^2$) [25]. NLL belongs to the family of *proper scoring rules*, a well-establish family of metrics for assessing model's calibration quality in the probabilistic forecast literature [28, 5].

On the other hand, given a model with predictive CDF $F(y|\mathbf{x})$, we compute ECE by first compute the model's $q\%$ predictive intervals $CI_q = \big[F^{-1}(\frac{q}{2}|\mathbf{x}_i), F^{-1}(1-\frac{q}{2}|\mathbf{x}_i)\big]$ for all observations $\{y_i, \mathbf{x}_i\}_{i=1}^N$, and estimate the empirical coverage probability as $\hat{q}(F) = \frac{1}{N}\sum_{i=1}^N \mathbb{I}\big(y_i \in CI_q(\mathbf{x}_i|F)\big)$. If the model's predictive distribution properly captures the empirical distribution of the data, the empirical coverage probability $\mathbb{P}(y \in CI_q|F)$ will be close to the nominal coverage probability q, regardless of the value of q. To this end, ECE summarizes the predictive model's quality in uncertainty quantification using the squared difference between q and $\mathbb{P}(y \in CI_q|F)$ for $q \in [0,1)$,

$$ECE = \int_0^1 ||q - \mathbb{P}(y \in CI_q|F)||_2^2 \, dq \tag{15}$$

We compare the performance of BNE with four other commonly used ensemble algorithms, summarized in Table 1. These are (1) the Bayesian stacking (**stack**), which aggregates models using non-adaptive weights but calibrates the predictive distribution using leave-one-out cross validation [87]; (2) the generalized additive model ensemble (**GAM**), which combines the base-model predictions in a regression model that includes additive, linear terms for each set of predictions and an extra smoothing spline term as a function of the feature space to mitigate any systematic bias [85]; (3) adaptive Bayesian model averaging (**BMA**), which aggregates models using spatially adaptive weights parameterized using Gaussian processes; and (4) Bayesian adaptive ensemble (**BAE**), which mitigates the predictive bias in BMA using an additive Gaussian process [60]. The BMA and BAE are reduced versions of BNE, where BMA corresponds

to a BNE model with no bias correction and no distribution calibration and BAE corresponds to a BNE model with no distribution calibration.

We compare the performance of these competing models when applied to both time-series and spatial data having complex mean structure and heterogeneous variability. For the time-series data, we consider a nonlinear time series $f = sin(x)$, with $x \in (-5\pi, 5\pi)$. We consider a data distribution of $y \sim logNormal(f(x), \sigma(x))$ with $\sigma(x) = exp(cos(x) - 1)$ so the it exhibits heterogeneous variance and skewness depending on the location of the input space (see Figure 3). (**jzl:** UPDATE Figure 3) For the spatial scenario, we use the multimodal Mishra's Bird function $f(x_1, x_2) = cos(4x_1) * exp\big((1 - sin(4x_2))^2\big) + sin(4x_2) * exp\big((1 - cos(4x_1))^2\big) + (x_1 - x_2)^2$ for $x_1, x_2 \in (-\pi, \pi)$. This data generating function, common in the machine learning literature, is characterized by a slow-varying global trend but with sharp local fluctuations [56] (Figure 4). We consider heterogeneous log-normal distribution $y \sim\sim N(f(x_1, x_2), \sigma^2(x_1, x_2))$, where the variance is the modified Rosenbrock function $\sigma(x_1, x_2) = 4 * (x_2 - x_1)^2 + (1 - x_1)^2$ so the data distribution exhibits higher variance and skewness especially at the margin of input space.

To prepare the base models for the ensemble methods, we specify a library of base models $\mathscr{F} = \{\hat{f}_k\}_{k=1}^K$, where each $\hat{f}_k$ is pre-trained on simulated data of sample size 50 (for time-series experiments) or 250 (for spatial experiments). In this work, we choose each f_k's to be kernel ridge regression models with the following kernel functions: (1) a polynomial kernel with degree 3 $k(x_1, x_2) = (1 + x_1^\top x_2)^3$, (2) an arc-cosine kernel $k(x_1, x_2) = 1 - \frac{1}{\pi} arccos(\frac{x_1^\top x_2}{|x_1||x_2|})$ whose model space is equivalent to that of a 1-layer neural network with infinite-number of neurons [8], an (3) exponential kernel $k(x_1, x_2) = exp(-|x_1 - x_2|/l)$ whose model space represents the space of continuous (and possibly non-differentiable) functions, and (4) Matérn $\frac{5}{2}$ kernel whose model space represents the space of continuous functions that are at least twice differentiable [64]. As a result, the library $\mathscr{F} = \{f_k\}_{k=1}^3$ as a whole encodes a diverse set of assumptions about the data generating function $f(\mathbf{x})$, but none of the individual models f_k can predict $f(\mathbf{x})$ universally well across the input space $\mathscr{X}$, especially in without training data (see Figure 3-4). Then, using $\mathscr{F}$, we fit each ensemble method in Table 1 on a simulated data of sample size $n \in (50, 1000)$, and we compute each model's operating characteristics with respect to prediction (root mean squared error (RMSE)) and uncertainty quantification (NLL and ECE) on a separate validation dataset with sample size 1000. We repeat the simulation for each model-sample size combination 100 times.

(**jzl:** TODO: To update Figures) Figure 5 illustrates the behavior of the predictive distribution of each ensemble method for one realization of the time-series case in which the noise process is complex an the base models are misspecified. Because **stack** and **BMA** generate predictions as a convex combination of the base model predictions, they produced a biased estimate for some local regions due to the systematic bias shared among the base models. **GAM** improved upon **stack** and **BMA** by nonparametrically correcting the base models' systematic bias using a smoothing spline. However, restricted by its Gaussian assumption, **GAM** produced a set of inflexible predictive distributions with constant variance. These distributions failed to capture the skewness and heterogeneity in the data distributions, leading to suboptimal uncertainty quantification. **BAE** improved upon **GAM** by estimating its predictive distribution adaptively as a flexible mixture of Gaussians, thereby yielding a set of predictive quantiles that accommodates the heterogeneity in the residual error variance across the feature space, but fails to capture

the skewness in the data distribution. Finally, **BNE** improved upon the performance of **BAE** by calibrating its predictive distribution toward the empirical distribution of the data, yielding a predictive distribution that properly captures both the heterogeneity and skewness in the data distribution. This flexibility led to predictive quantiles having the correct coverage even under complex data generating mechanisms.

Tables 2 and 3 reports the operating characteristics (MSE, NLL, ECE, and coverage probability of 95% credible intervals (CI)) for each ensemble method in the time-series and spatial settings, respectively. An important goal of this simulation study is to understand if the quality of the ensemble model's predictive uncertainty remains robust even in regions that are outside training data. To this end, we test the model's uncertainty performance against a input space that is larger than that covered by the training data. That is, a test time window of $(-5\pi, 5\pi)$ versus a training time window of $(-3\pi, 3\pi)$ in the time-series experiments, and a test region of $(-1.25\pi, 1.25\pi)^2$ versus a training region of $(-\pi, \pi)^2$ in the spatial experiments. Ideally, we expect the ensemble model to provide not only accurate prediction for test samples from the training region (as measured by MSE), and also calibrated uncertainty quantification across the whole test region even if it is not covered by the training data (as measured by NLL, ECE and coverage probability of 95% CIs).

As shown, the relative performances of the various ensemble models are largely consistent across both scenarios. Specifically, as the sample size increases, both the RMSE and ECE for **stack** and **BMA** are higher than that of the other approaches, due to their inability to correct for bias in the base models. **GAM** and **BAE** improved upon the previous two methods by adapting the mean prediction to the data. However, the inflexibility in the distributional assumptions made by these approaches prevented them from capturing the heterogeneity and skewness in the error distribution, producing sub-optimal uncertainty estimates (both in terms of the proper scoring rule, as measured by NLL, and coverage of predictive interval, as measure by ECE) even in large samples. On the other hand, **BNE** improved with increasing sample size due to the flexibility in its distributional assumptions.

To further understand the behavior of these models, we decompose the NLL score into its calibration and sharpness components (i.e., $\frac{(y-\mu(x))^2}{\sigma(x)^2}$ and $log\sigma$, respectively), and also measure the coverage probability of model's 95% predictive credible intervals, both in the training region and in the test region. Tables 5-6 in the Supplementary Material shows the results. As shown, comparing to models with inflexible distribution assumption, the **BNE** model achieves both better calibration and on average narrower credible intervals (i.e., better sharpness). At the same time, the **BNE**'s 95% credible intervals better maintains the nomial coverage, both in the training region and across the whole test region. The advantage is especially prominent in the case of small samples ($n = 10$) and in the case of complex spatial heterogeneity (spatial experiments). This shows that the **BNE** model is not naively improving model uncertainty by widening its credible intervals everywhere, and the **BNE**'s flexible distributional model does not lead to overfitting, but instead leads it to better capture complex patterns in the training data, and better expresses epistemic uncertainty when out of the training regions.

6 Spatial ensemble of $PM_{2.5}$ levels in Eastern New England, USA

(**jzl:** yet to update this section, please skip.)

Here, we use BNE to integrate information on $PM_{2.5}$ levels from the three distinct $PM_{2.5}$ exposure models introduced in Section 1.2 and produce a single set of predicted concentrations for average $PM_{2.5}$ levels in 2011. Our goals are to (1) improve prediction accuracy relative to any one set of predictions; (2) understand the driving factors behind the ensemble system's uncertainty; and (3) understand spatial locations where the available exposure models systematically underperform. We implement our ensemble framework described in Sections 2 and 4 on the three exposure models' out-of-sample prediction for 43 monitors, along with the other four ensemble approaches that were also considered in the simulation studies. For all Bayesian methods, we perform posterior inference using Hamiltonian Monte Carlo (HMC) with sample size 10^4 after using a burn-in of 1000. We present the results in Table 4. The posterior estimates from the BNE model consistently returned the lowest RMSE, indicating improved quality in both predictive accuracy and uncertainty calibration when compared to that of the existing approaches considered.

The bottom panel of Figure 1 presents the posterior estimates of the spatially-adaptive ensemble weights for the Kloog, Di, and van Donkelaar models from the BNE fit. Finally, BNE fit in general assigns low weights to the van Donkelaar model.

These results are the first to quantitatively compare the predictive accuracy of these three models in this region. The Di model has high predictive accuracy in an urban area with a dense monitoring network. A potential explanation for is that because this model uses data from the entire country it is driven primarily by monitoring data in similar urban settings. The Kloog model, on the other hand, may do better in more rural, low monitoring settings because it was trained specifically on this region, as opposed to nationally, and so may be more specific to this region. In contrast, the van Donkelaar model was trained at an even broader area (the entire North American region) and this appears to have lowered the predictive accuracy of this model relative to the others.

Figure 6 visualizes the posterior predictive distribution for $PM_{2.5}$ levels over the study region from the BNE fit. Specifically, Figures 6a–6b show BNE's posterior predictive mean and predictive uncertainty (posterior predictive variance), respectively. Figures 6c-6d show the decomposition of the predictive uncertainty into that due to model uncertainty (i.e. the posterior variance associated with the ensemble weights ω) and that due to model prediction (i.e. the posterior variance for δ and G), respectively. In all figures, the black circles indicate the locations of the air pollution monitors. 6c shows higher levels of uncertainty southwest of Boston and also in regions that are farther away from this urban area. Comparing these uncertainty estimates to the model-specific predictions in Figure 1, we notice that these are the regions where the Kloog and Di models tend to disagree. Figure 6d, furthermore, shows that the lack of monitoring data in this region also plays a role in the increased uncertainty in the western part of our study region. To our knowledge, this is the first comprehensive analysis that can assess the relative performance of multiple prediction models, how these vary across an area, and how data sparsity and disagreements among model outputs contribute to the overall uncertainty associated with the ensemble estimates.

7 Conclusion

In this work, we have presented a principled Bayesian nonparametric approach for combining deterministic model predictions to construct a *spatially adaptive* ensemble that produces *calibrated* predictive uncertainty. To illustrate the practical effectiveness of the proposed approach, we compared via simulation the performance of the BNE model with that of several existing ensemble methods. This comparison showed clear improvements in both out-of-sample predictive accuracy and in uncertainty quantification. Finally, we applied the proposed approach to integrate three air pollution prediction models for predicting annual $PM_{2.5}$ concentrations in eastern New England. The resulting predictions outperformed those of the existing ensemble approaches in terms of out-of-sample RMSE.

By producing a set of exposure predictions with rigorously quantified uncertainty, the proposed BNE model can be used to provide calibrated exposure estimates for downstream analyses. For instance, there is currently great interest in identifying characteristics of locations (e.g. urban, rural, region of the country) and times (e.g. seasons or years) most associated with high uncertainty of of air pollution estimates, and how such varying prediction error may impact downstream epidemiologic health effects analyses. A few recent efforts have sought to develop measurement-error corrections applicable to such spatio-temporally varying measurement error [2]), and the uncertainty estimates obtained from our proposed framework can serve as inputs into such approaches. Moreoever, reporting the estimated model weights and uncertainty estimates from this approach back to teams developing the original prediction models can inform efforts to improve these models. For instance, identification of high uncertainty areas can help inform future monitoring campaigns or suggest additional model inputs that could improve model performance.

The application we considered focused on the integration of *spatial* prediction models for annual $PM_{2.5}$ averages. In practice, the such integration of multiple *spatio-temporal* air pollution prediction models is also of interest in many applications. Adapting BNE to the spatio-temporal setting is conceptually straightforward. We can incorporate time as an additional input to the ensemble weight $\omega_k(\mathbf{x})$. However, this spatio-temporal setting raises a few novel challenges that require careful treatment for the model to remain effective. The first is additional model complexity introduced by the temporal dimension. For spatio-temporal data, the ensemble model needs to be sufficiently flexible to handle temporal trends in air pollution concentrations, such as seasonal variability, that are distinct from its spatial variability. In the context of BNE, these two objectives can be achieved by designing a temporal kernel function $k_{time}(t,t')$ that models the temporal dimension using smoothing temporal basis functions similar to those used in classic spatio-temporal models for air pollution [38, 49, 71]. Then, the spatio-temporal dependency can be modelled by an interaction kernel $k_{space\times time} = k_{space} \otimes k_{time}$ that is constructed as a tensor product, and is regularized by certain sparse-inducing priors (e.g., horseshoe), leading to a final kernel in the form of $k = k_{space} + k_{time} + k_{space\times time}$ [31, 46]. We leave the exploration for the optimal kernel design for modeling spatio-temporal interactions to future work.

The distribution model developed in this work (**HEX-GP**, Section 2.1) is tailored for the air pollution applications, where the data $y|\mathbf{x}$ exhibits positive skewness and heavy tail. To this end, our focus was to illustrate the importance of distribution calibration for the model's ability

in uncertainty quantification, and have presented a practical approach that induces sufficient flexibility in capturing the non-Gaussian characteristics in the data, while maintaining analytical and computational tractability. However, it should also be recognized that the model space of **HEX-GP** does not contain all the possible distribution patterns, e.g., negative skewness or extreme kurtosis. For applications where modeling such phenomenon is important, we note that it is straightforward to extend the recipe in (7) to further complexity. For example, the negative skewness can be modeled by introducing double-exponential modification to the mean component $\mu(\mathbf{x})$ [65], and the extreme kurtosis can be modeled by introducing a inverse Gamma prior to $s(\mathbf{x})$. However, this added flexibility invariably introduces additional (nonparametric) parameters to the model, incurring a trade-off between the model's flexibility and its statistical / computational efficiency. To this end, we recommend practitioners to take a step-wise approach to model inference (similar to Section 4) to ensure the added complexity brings meaningful improvement to model performance.

Overall, we anticipate the benefits of adaptive weights that maximize predictive accuracy and, thus, minimize exposure measurement error to be widely applicable in a variety of settings in which space-time prediction is of interest. The approach yields rigorous uncertainty quantification that can be used to answer multiple relevant scientific questions, likely in multiple fields. We have found that this approach more effectively characterizes the sources of uncertainty associated with prediction of fine particulate matter concentrations in Eastern New England.

References

[1] J. Abernethy, P. L. Bartlett, and E. Hazan. Blackwell Approachability and No-Regret Learning are Equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 27–46, Dec. 2011.

[2] S. E. Alexeeff, R. J. Carroll, and B. Coull. Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures. *Biostatistics*, 17(2):377–389, 2016.

[3] P. D. Arendt, D. W. Apley, and W. Chen. Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability. *Journal of Mechanical Design*, 134(10):100908–100908–12, Sept. 2012.

[4] J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. sgmcmc: An r package for stochastic gradient markov chain monte carlo. *Journal of Statistical Software*, 91(3):1âĂŞ27, 2019.

[5] J. M. Bernardo. Expected information as expected utility. *the Annals of Statistics*, pages 686–690, 1979.

[6] J. Brynjarsdóttir and A. OÊijHagan. Learning about physical parameters: The importance of model discrepancy. *Inverse problems*, 30(11):114007, 2014.

[7] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge ; New York, Mar. 2006.

[8] Y. Cho and L. K. Saul. Kernel Methods for Deep Learning. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 342–350. Curran Associates, Inc., 2009.

[9] K. Choromanski, M. Rowland, T. Sarlos, V. Sindhwani, R. Turner, and A. Weller. The Geometry of Random Features. In *International Conference on Artificial Intelligence and Statistics*, pages 1–9, Mar. 2018.

[10] M. Clydec and E. S. Iversen. *Bayesian model averaging in the M-open framework*. Oxford University Press, Jan. 2013.

[11] A. J. Cohen, M. Brauer, R. Burnett, H. R. Anderson, J. Frostad, K. Estep, K. Balakrishnan, B. Brunekreef, L. Dandona, R. Dandona, and others. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082):1907–1918, 2017.

[12] K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. Random Feature Expansions for Deep Gaussian Processes. In *International Conference on Machine Learning*, pages 884–893, July 2017.

[13] A. P. Dawid. The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

[14] A. Der Kiureghian and O. Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[15] Q. Di, H. Amini, L. Shi, I. Kloog, R. Silvern, J. Kelly, M. B. Sabath, C. Choirat, P. Koutrakis, A. Lyapustin, Y. Wang, L. J. Mickley, and J. Schwartz. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, 130:104909, Sept. 2019.

[16] Q. Di, L. Dai, Y. Wang, A. Zanobetti, C. Choirat, J. D. Schwartz, and F. Dominici. Association of Short-term Exposure to Air Pollution With Mortality in Older Adults. *JAMA*, 318(24):2446–2456, 2017.

[17] Q. Di, I. Kloog, P. Koutrakis, A. Lyapustin, Y. Wang, and J. Schwartz. Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States. *Environmental Science & Technology*, 50(9):4712–4721, May 2016.

[18] Q. Di, P. Koutrakis, and J. Schwartz. A hybrid prediction model for PM2.5 mass and components using a chemical transport model and land use regression. *Atmospheric Environment*, 131:390–399, Apr. 2016.

[19] Q. Di, Y. Wang, A. Zanobetti, Y. Wang, P. Koutrakis, C. Choirat, F. Dominici, and J. D. Schwartz. Air pollution and mortality in the Medicare population. *New England Journal of Medicine*, 376(26):2513–2522, 2017.

[20] K. S. Dorans, E. H. Wilker, W. Li, M. B. Rice, P. L. Ljungman, J. Schwartz, B. A. Coull, I. Kloog, P. Koutrakis, R. B. DâĂŹAgostino, J. M. Massaro, U. Hoffmann, C. J. OâĂŹDonnell, and M. A. Mittleman. Residential Proximity to Major Roads, Exposure to Fine Particulate Matter and Coronary Artery Calcium: The Framingham Heart Study. *Arteriosclerosis, thrombosis, and vascular biology*, 36(8):1679–1685, Aug. 2016.

[21] M. Eeftens, R. Beelen, K. de Hoogh, T. Bellander, G. Cesaroni, M. Cirach, C. Declercq, A. DÄŮdelÄŮ, E. Dons, A. de Nazelle, K. Dimakopoulou, K. Eriksen, G. Falq, P. Fischer, C. Galassi, R. GraÅ¿uleviÄ■ienÄŮ, J. Heinrich, B. Hoffmann, M. Jerrett, D. Keidel, M. Korek, T. Lanki, S. Lindley, C. Madsen, A. MÃűlter, G. NÃądor, M. Nieuwenhuijsen, M. Nonnemacher, X. Pedeli, O. Raaschou-Nielsen, E. Patelarou, U. Quass, A. Ranzi, C. Schindler, M. Stempfelet, E. Stephanou, D. Sugiri, M.-Y. Tsai, T. Yli-Tuomi, M. J. VarrÃş, D. Vienneau, S. v. Klot, K. Wolf, B. Brunekreef, and G. Hoek. Development of Land Use Regression models for PM(2.5), PM(2.5) absorbance, PM(10) and PM(coarse) in 20 European study areas. *Environmental Science & Technology*, 46(20):11195–11205, Oct. 2012.

[22] A. F. Fleisch, H. Luttmann-Gibson, W. Perng, S. L. Rifas-Shiman, B. A. Coull, I. Kloog, P. Koutrakis, J. D. Schwartz, A. Zanobetti, C. S. Mantzoros, M. W. Gillman, D. R. Gold, and E. Oken. Prenatal and early life exposure to traffic pollution and cardiometabolic health in childhood. *Pediatric Obesity*, 12(1):48–57, 2017.

[23] M. H. Forouzanfar, A. Afshin, L. T. Alexander, H. R. Anderson, Z. A. Bhutta, S. Biryukov, M. Brauer, R. Burnett, K. Cercy, F. J. Charlson, A. J. Cohen, L. Dandona, K. Estep, A. J. Ferrari, J. J. Frostad, N. Fullman, P. W. Gething, W. W. Godwin, M. Griswold, S. I. Hay, Y. Kinfu, H. H. Kyu, H. J. Larson, X. Liang, S. S. Lim, P. Y. Liu, A. D. Lopez, R. Lozano, L. Marczak, G. A. Mensah, A. H. Mokdad, M. Moradi-Lakeh, M. Naghavi, B. Neal, M. B. Reitsma, G. A. Roth, J. A. Salomon, P. J. Sur, T. Vos, J. A. Wagner, H. Wang, Y. Zhao, M. Zhou, G. M. Aasvang, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, A. M. Abdulle, S. F. Abera, B. Abraham, L. J. Abu-Raddad, G. Y. Abyu, A. O. Adebiyi, I. A. Adedeji, Z. Ademi, A. K. Adou, J. C. Adsuar, E. E. Agardh, A. Agarwal, A. Agrawal, A. A. Kiadaliri, O. N. Ajala, T. F. Akinyemiju, Z. Al-Aly, K. Alam, N. K. M. Alam, S. F. Aldhahri, R. W. Aldridge, Z. A. Alemu, R. Ali, A. Alkerwi, F. Alla, P. Allebeck, U. Alsharif, K. A. Altirkawi, E. A. Martin, N. Alvis-Guzman, A. T. Amare, A. Amberbir, A. K. Amegah, H. Amini, W. Ammar, S. M. Amrock, H. H. Andersen, B. O. Anderson, C. A. T. Antonio, P. Anwari, J. ÃĎrnlÃűv, A. Artaman, H. Asayesh, R. J. Asghar, R. Assadi, S. Atique, E. F. G. A. Avokpaho, A. Awasthi, B. P. A. Quintanilla, P. Azzopardi, U. Bacha, A. Badawi, M. C. Bahit, K. Balakrishnan, A. Barac, R. M. Barber, S. L. Barker-Collo, T. BÃd'rnighausen, S. Barquera, L. Barregard, L. H. Barrero, S. Basu, C. Batis, S. Bazargan-Hejazi, J. Beardsley, N. Bedi, E. Beghi, B. Bell, M. L. Bell, A. K. Bello, D. A. Bennett, I. M. Bensenor, A. Berhane, E. BernabÃl', B. D. Betsu, A. S. Beyene, N. Bhala, A. Bhansali, S. Bhatt, S. Biadgilign, B. Bikbov, D. Bisanzio, E. Bjertness, J. D. Blore, R. Borschmann, S. Boufous, R. R. A. Bourne, M. Brainin, A. Brazinova, N. J. K. Breitborde, H. Brenner, D. M. Broday, T. S. Brugha, B. Brunekreef, Z. A. Butt, L. E. Cahill, B. Calabria, I. R. Campos-Nonato, R. CÃąrdenas, D. O. Carpenter, J. J. Carrero, D. C. Casey, C. A. CastaÃśeda-Orjuela, J. C. Rivas, R. E. Castro, F. CatalÃą-LÃşpez, J.-C. Chang, P. P.-C. Chiang, M. Chibalabala, O. Chimed-Ochir, V. H. Chisumpa, A. A. Chitheer, J.-Y. J. Choi, H. Christensen, D. J. Christopher, L. G. Ciobanu, M. M. Coates, S. M. Colquhoun, A. G. C. Manzano, L. T. Cooper, K. Cooperrider, L. Cornaby, M. Cortinovis, J. A. Crump, L. Cuevas-Nasu, A. Damasceno, R. Dandona, S. C. Darby, P. I. Dargan, J. d. Neves, A. C. Davis, K. Davletov, E. F. d. Castro, V. D. l. Cruz-GÃşngora, D. D. Leo, L. Degenhardt, L. C. D. Gobbo, B. d. Pozo-Cruz, R. P. Dellavalle, A. Deribew, D. C. D. Jarlais, S. D. Dharmaratne, P. K. Dhillon, C. Diaz-TornÃl', D. Dicker, E. L. Ding, E. R. Dorsey, K. E. Doyle, T. R. Driscoll, L. Duan, M. Dubey, B. B. Duncan, I. Elyazar, A. Y. Endries, S. P. Ermakov, H. E. Erskine, B. Eshrati, A. Esteghamati, S. Fahimi, E. J. A. Faraon, T. A. Farid, C. S. e. S. Farinha, A. Faro, M. S. Farvid, F. Farzadfar, V. L. Feigin, S.-M. Fereshtehnejad, J. G. Fernandes, F. Fischer, J. R. A. Fitchett, T. Fleming, N. Foigt, K. Foreman, F. G. R. Fowkes, R. C. Franklin, T. FÃijrst, N. D. Futran, E. Gakidou, A. L. Garcia-Basteiro, T. T. Gebrehiwot, A. T. Gebremedhin, J. M. Geleijnse, B. D. Gessner, A. Z. Giref, M. Giroud, M. D. Gishu, G. Giussani, S. Goenka, M. C. Gomez-Cabrera, H. Gomez-Dantes, P. Gona, A. Goodridge, S. V. Gopalani, C. C. Gotay, A. Goto, H. N. Gouda, H. C. Gugnani, F. Guillemin, Y. Guo, R. Gupta, R. Gupta, R. A. GutiÃl'rrez, J. A. Haagsma, N. Hafezi-Nejad, D. Haile, G. B. Hailu, Y. A. Halasa, R. R. Hamadeh, S. Hamidi, A. J. Handal, G. J. Hankey, Y. Hao, H. L. Harb, S. Harikrishnan, J. M. Haro, M. S. Hassanvand, T. A. Hassen, R. Havmoeller, I. B. Heredia-Pi, N. F. HernÃąndez-Llanes, P. Heydarpour, H. W.

Hoek, H. J. Hoffman, M. Horino, N. Horita, H. D. Hosgood, D. G. Hoy, M. Hsairi, A. S. Htet, G. Hu, J. J. Huang, A. Husseini, S. J. Hutchings, I. Huybrechts, K. M. Iburg, B. T. Idrisov, B. V. Ileanu, M. Inoue, T. A. Jacobs, K. H. Jacobsen, N. Jahanmehr, M. B. Jakovljevic, H. A. F. M. Jansen, S. K. Jassal, M. Javanbakht, S. P. Jayaraman, A. U. Jayatilleke, S. H. Jee, P. Jeemon, V. Jha, Y. Jiang, T. Jibat, Y. Jin, C. O. Johnson, J. B. Jonas, Z. Kabir, Y. Kalkonde, R. Kamal, H. Kan, A. Karch, C. K. Karema, C. Karimkhani, A. Kasaeian, A. Kaul, N. Kawakami, D. S. Kazi, P. N. Keiyoro, L. Kemmer, A. H. Kemp, A. P. Kengne, A. Keren, C. N. Kesavachandran, Y. S. Khader, A. R. Khan, E. A. Khan, G. Khan, Y.-H. Khang, S. Khatibzadeh, S. Khera, T. A. M. Khoja, J. Khubchandani, C. Kieling, C.-i. Kim, D. Kim, R. W. Kimokoti, N. Kissoon, M. Kivipelto, L. D. Knibbs, Y. Kokubo, J. A. Kopec, P. A. Koul, A. Koyanagi, M. Kravchenko, H. Kromhout, H. Krueger, T. Ku, B. K. Defo, R. S. Kuchenbecker, B. K. Bicer, E. J. Kuipers, G. A. Kumar, G. F. Kwan, D. K. Lal, R. Lalloo, T. Lallukka, Q. Lan, A. Larsson, A. A. Latif, A. E. B. Lawrynowicz, J. L. Leasher, J. Leigh, J. Leung, M. Levi, X. Li, Y. Li, J. Liang, S. Liu, B. K. Lloyd, G. Logroscino, P. A. Lotufo, R. Lunevicius, M. MacIntyre, M. Mahdavi, M. Majdan, A. Majeed, R. Malekzadeh, D. C. Malta, W. A. A. Manamo, C. C. Mapoma, W. Marcenes, R. V. Martin, J. Martinez-Raga, F. Masiye, K. Matsushita, R. Matzopoulos, B. M. Mayosi, J. J. McGrath, M. McKee, P. A. Meaney, C. Medina, A. Mehari, F. Mejia-Rodriguez, A. B. Mekonnen, Y. A. Melaku, Z. A. Memish, W. Mendoza, G. B. M. Mensink, A. Meretoja, T. J. Meretoja, Y. M. Mesfin, F. A. Mhimbira, A. Millear, T. R. Miller, E. J. Mills, M. Mirarefin, A. Misganaw, C. N. Mock, A. Mohammadi, S. Mohammed, G. L. D. Mola, L. Monasta, J. C. M. Hernandez, M. Montico, L. Morawska, R. Mori, D. Mozaffarian, U. O. Mueller, E. Mullany, J. E. Mumford, G. V. S. Murthy, J. B. Nachega, A. Naheed, V. Nangia, N. Nassiri, J. N. Newton, M. Ng, Q. L. Nguyen, M. I. Nisar, P. M. N. Pete, O. F. Norheim, R. E. Norman, B. Norrving, L. Nyakarahuka, C. M. Obermeyer, F. A. Ogbo, I.-H. Oh, O. Oladimeji, P. R. Olivares, H. Olsen, B. O. Olusanya, J. O. Olusanya, J. N. Opio, E. Oren, R. Orozco, A. Ortiz, E. Ota, M. Pa, A. Pana, E.-K. Park, C. D. Parry, M. Parsaeian, T. Patel, A. J. P. Caicedo, S. T. Patil, S. B. Patten, G. C. Patton, N. Pearce, D. M. Pereira, N. Perico, K. Pesudovs, M. Petzold, M. R. Phillips, F. B. Piel, J. D. Pillay, D. Plass, S. Polinder, C. D. Pond, C. A. Pope, D. Pope, S. Popova, R. G. Poulton, F. Pourmalek, N. M. Prasad, M. Qorbani, R. H. S. Rabiee, A. Radfar, A. Rafay, V. Rahimi-Movaghar, M. Rahman, M. H. U. Rahman, S. U. Rahman, R. K. Rai, S. Rajsic, M. Raju, U. Ram, S. M. Rana, K. Ranganathan, P. Rao, C. A. R. GarcÃ■a, A. H. Refaat, C. D. Rehm, J. Rehm, N. Reinig, G. Remuzzi, S. Resnikoff, A. L. Ribeiro, J. A. Rivera, H. S. Roba, A. Rodriguez, S. Rodriguez-Ramirez, D. Rojas-Rueda, Y. Roman, L. Ronfani, G. Roshandel, D. Rothenbacher, A. Roy, M. M. Saleh, J. R. Sanabria, L. Sanchez-Riera, M. D. Sanchez-NiÃśo, T. G. SÃąnchez-Pimienta, L. Sandar, D. F. Santomauro, I. S. Santos, R. Sarmiento-Suarez, B. Sartorius, M. Satpathy, M. Savic, M. Sawhney, J. Schmidhuber, M. I. Schmidt, I. J. C. Schneider, B. SchÃűttker, A. E. Schutte, D. C. Schwebel, J. G. Scott, S. Seedat, S. G. Sepanlou, E. E. Servan-Mori, G. Shaddick, A. Shaheen, S. Shahraz, M. A. Shaikh, T. S. Levy, R. Sharma, J. She, S. Sheikhbahaei, J. Shen, K. N. Sheth, P. Shi, K. Shibuya, M. Shigematsu, M.-J. Shin, R. Shiri, K. Shishani, I. Shiue, M. G. Shrime, I. D. Sigfusdottir, D. A. S. Silva, D. G. A. Silveira, J. I. Silverberg, E. P. Simard, S. Sindi,

A. Singh, J. A. Singh, P. K. Singh, E. L. Slepak, M. Soljak, S. Soneji, R. J. D. Sorensen, L. A. Sposato, C. T. Sreeramareddy, V. Stathopoulou, N. Steckling, N. Steel, D. J. Stein, M. B. Stein, H. StÃűckl, S. Stranges, K. Stroumpoulis, B. F. Sunguya, S. Swaminathan, B. L. Sykes, C. E. I. Szoeke, R. TabarÃl's-Seisdedos, K. Takahashi, R. T. Talongwa, N. Tandon, D. Tanne, M. Tavakkoli, B. W. Taye, H. R. Taylor, B. A. Tedla, W. M. Tefera, T. K. Tegegne, D. Y. Tekle, A. S. Terkawi, J. S. Thakur, B. A. Thomas, M. L. Thomas, A. J. Thomson, A. L. Thorne-Lyman, A. G. Thrift, G. D. Thurston, T. Tillmann, R. Tobe-Gai, M. Tobollik, R. Topor-Madry, F. Topouzis, J. A. Towbin, B. X. Tran, Z. T. Dimbuene, N. Tsilimparis, A. K. Tura, E. M. Tuzcu, S. Tyrovolas, K. N. Ukwaja, E. A. Undurraga, C. J. Uneke, O. A. Uthman, A. v. Donkelaar, J. v. Os, Y. Y. Varakin, T. Vasankari, J. L. Veerman, N. Venketasubramanian, F. S. Violante, S. E. Vollset, G. R. Wagner, S. G. Waller, J. L. Wang, L. Wang, Y. Wang, S. Weichenthal, E. Weiderpass, R. G. Weintraub, A. Werdecker, R. Westerman, H. A. Whiteford, T. Wijeratne, C. S. Wiysonge, C. D. A. Wolfe, S. Won, A. D. Woolf, M. Wubshet, D. Xavier, G. Xu, A. K. Yadav, B. Yakob, A. Z. Yalew, Y. Yano, M. Yaseri, P. Ye, P. Yip, N. Yonemoto, S.-J. Yoon, M. Z. Younis, C. Yu, Z. Zaidi, M. E. S. Zaki, J. Zhu, B. Zipkin, S. Zodpey, L. J. Zuhlke, and C. J. L. Murray. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990âĂŞ2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053):1659–1724, Oct. 2016.

[24] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

[25] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, Apr. 2007.

[26] T. Gneiting and A. E. Raftery. Weather Forecasting with Ensemble Methods. *Science*, 310(5746):248–249, Oct. 2005.

[27] T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, Mar. 2007.

[28] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114, 1952.

[29] E. Grushka. Characterization of exponentially modified gaussian peaks in chromatography. *Analytical chemistry*, 44 11:1733–8, 1972.

[30] A. Gryparis, C. J. Paciorek, A. Zeka, J. Schwartz, and B. A. Coull. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10(2):258–274, 2008.

[31] C. Gu. *Smoothing Spline ANOVA Models*. Springer Science & Business Media, Jan. 2013. Google-Books-ID: 5VxGAAAAQBAJ.

[32] M. D. Hoffman and A. Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv:1111.4246 [cs, stat]*, Nov. 2011. arXiv: 1111.4246.

[33] X. Hu, J. H. Belle, X. Meng, A. Wildani, L. A. Waller, M. J. Strickland, and Y. Liu. Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environmental Science & Technology*, 51(12):6936–6944, June 2017.

[34] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[35] A. Jara and T. E. Hanson. A class of mixtures of dependent tail-free processes. *Biometrika*, 98(3):553–566, Sept. 2011.

[36] I. Jhun, J. Kim, B. Cho, D. R. Gold, J. Schwartz, B. A. Coull, A. Zanobetti, M. B. Rice, M. A. Mittleman, E. Garshick, P. Vokonas, M.-A. Bind, E. H. Wilker, F. Dominici, H. Suh, and P. Koutrakis. Synthesis of Harvard Environmental Protection Agency (EPA) Center studies on traffic-related particulate pollution and cardiovascular outcomes in the Greater Boston Area. *Journal of the Air & Waste Management Association*, 69(8):900–917, Aug. 2019.

[37] X. Jin, A. M. Fiore, K. Civerolo, J. Bi, Y. Liu, A. Van Donkelaar, R. V. Martin, M. Al-Hamdan, Y. Zhang, T. Z. Insaf, et al. Comparison of multiple pm2. 5 exposure products for estimating health benefits of emission controls over new york state, usa. *Environmental Research Letters*, 14(8):084023, 2019.

[38] Keller Joshua P., Olives Casey, Kim Sun-Young, Sheppard Lianne, Sampson Paul D., Szpiro Adam A., Oron Assaf P., LindstrÃűm Johan, Vedal Sverre, and Kaufman Joel D. A Unified Spatiotemporal Modeling Approach for Predicting Concentrations of Multiple Air Pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution. *Environmental Health Perspectives*, 123(4):301–309, Apr. 2015.

[39] J. T. Kelly, C. Jang, B. Timin, Q. Di, J. Schwartz, Y. Liu, A. van Donkelaar, R. V. Martin, V. Berrocal, and M. L. Bell. Examining pm2. 5 concentrations and exposure using multiple models. *Environmental Research*, page 110432, 2020.

[40] S.-Y. Kim, C. Olives, L. Sheppard, P. D. Sampson, T. V. Larson, J. P. Keller, and J. D. Kaufman. Historical Prediction Modeling Approach for Estimating Long-Term Concentrations of PM2.5 in Cohort Studies before the 1999 Implementation of Widespread Monitoring. *Environmental Health Perspectives*, 125(1):38–46, 2017.

[41] A. D. Kiureghian and O. Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, Mar. 2009.

[42] I. Kloog, A. A. Chudnovsky, A. C. Just, F. Nordio, P. Koutrakis, B. A. Coull, A. Lyapustin, Y. Wang, and J. Schwartz. A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmospheric Environment*, 95:581–590, Oct. 2014.

[43] I. Kloog, P. Koutrakis, B. A. Coull, H. J. Lee, and J. Schwartz. Assessing temporally and spatially resolved $PM_{2.5}$ exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric environment*, 45(35):6267–6275, 2011.

[44] I. Kloog, F. Nordio, B. A. Coull, and J. Schwartz. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal $PM_{2.5}$ exposures in the Mid-Atlantic states. *Environmental science & technology*, 46(21):11913–11921, 2012.

[45] V. Kuleshov and S. Ermon. Estimating Uncertainty Online Against an Adversary. In *Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence*, 2017.

[46] H. Li and D. Pati. Variable selection using shrinkage priors. *Computational Statistics & Data Analysis*, 107:107–119, Mar. 2017.

[47] L. Li, F. Lurmann, R. Habre, R. Urman, E. Rappaport, B. Ritz, J.-C. Chen, F. D. Gilliland, and J. Wu. Constrained Mixed-Effect Models with Ensemble Learning for Prediction of Nitrogen Oxides Concentrations at High Spatiotemporal Resolution. *Environmental Science & Technology*, 51(17):9920–9929, Sept. 2017.

[48] L. Li, J. Zhang, W. Qiu, J. Wang, and Y. Fang. An Ensemble Spatiotemporal Model for Predicting PM2.5 Concentrations. *International Journal of Environmental Research and Public Health*, 14(5), May 2017.

[49] J. LindstrÃűm, A. A. Szpiro, P. D. Sampson, A. P. Oron, M. Richards, T. V. Larson, and L. Sheppard. A Flexible Spatio-Temporal Model for Air Pollution with Spatial and Spatio-Temporal Covariates. *Environmental and ecological statistics*, 21(3):411–433, Sept. 2014.

[50] F. Liu, X. Huang, Y. Chen, and J. A. K. Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[51] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

[52] H. G. Matthies. Quantifying uncertainty: modern computational representation of probability and applications. In *Extreme man-made and natural hazards in dynamics of structures*, pages 105–135. Springer, 2007.

[53] L. A. McGuinn, C. Ward-Caviness, L. M. Neas, A. Schneider, Q. Di, A. Chudnovsky, J. Schwartz, P. Koutrakis, A. G. Russell, V. Garcia, and others. Fine particulate matter and cardiovascular disease: Comparison of assessment methods for long-term exposure. *Environmental Research*, 159:16–23, 2017.

[54] S. Mei, T. Misiakiewicz, and A. Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pages 3351–3418. PMLR, 2021.

[55] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

[56] S. K. Mishra. Some New Test Functions for Global Optimization and Performance of Repulsive Particle Swarm Method. SSRN Scholarly Paper ID 926132, Social Science Research Network, Rochester, NY, Aug. 2006.

[57] P. Muldowney, K. Ostaszewski, and W. Wojdowski. The Darth Vader rule. *Tatra Mountains Mathematical Publications*, 52(1), 2012.

[58] A. H. Murphy and E. S. Epstein. Verification of Probabilistic Predictions: A Brief Review. *Journal of Applied Meteorology*, 6(5):748–755, Oct. 1967.

[59] N. Murray, H. H. Chang, H. Holmes, and Y. Liu. Combining Satellite Imagery and Numerical Model Simulation to Estimate Ambient Air Pollution: An Ensemble Averaging Approach. *arXiv:1802.03077 [stat]*, Feb. 2018. arXiv: 1802.03077.

[60] N. L. Murray, H. A. Holmes, Y. Liu, and H. H. Chang. A Bayesian ensemble approach to combine PM2.5 estimates from statistical models using satellite imagery and numerical model simulation. *Environmental Research*, 178:108601, Nov. 2019.

[61] D. Pati, D. B. Dunson, and S. T. Tokdar. Posterior consistency in conditional distribution estimation. *Journal of multivariate analysis*, 116:456–472, Apr. 2013.

[62] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5):1155–1174, May 2005.

[63] A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.

[64] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. University Press Group Limited, Jan. 2006. Google-Books-ID: vWtwQgAACAAJ.

[65] E. Sager and O. A. Timoshenko. The double emg distribution and trade elasticities. *Canadian Journal of Economics/Revue canadienne d'économique*, 52(4):1523–1557, 2019.

[66] T. Seidenfeld. Calibration, Coherence, and Scoring Rules. *Philosophy of Science*, 52(2):274–294, 1985.

[67] G. Shaddick, M. L. Thomas, A. Jobling, M. Brauer, A. van Donkelaar, R. Burnett, H. Chang, A. Cohen, R. Van Dingenen, C. Dora, S. Gumy, Y. Liu, R. Martin, L. A. Waller, J. West, J. V. Zidek, and A. PrÃijss-UstÃijn. Data Integration Model for Air Quality: A Hierarchical Approach to the Global Estimation of Exposures to Ambient Air Pollution. *arXiv:1609.00141 [stat]*, Sept. 2016. arXiv: 1609.00141.

[68] R. L. Smith, C. Tebaldi, D. Nychka, and L. O. Mearns. Bayesian Modeling of Uncertainty in Ensembles of Climate Models. *Journal of the American Statistical Association*, 104(485):97–116, Mar. 2009.

[69] T. J. Sullivan. *Introduction to Uncertainty Quantification*. Texts in Applied Mathematics. Springer International Publishing, 2015.

[70] A. A. Szpiro, C. J. Paciorek, and L. Sheppard. Does More Accurate Exposure Prediction Necessarily Improve Health Effect Estimates? *Epidemiology (Cambridge, Mass.)*, 22(5):680–685, Sept. 2011.

[71] A. A. Szpiro, P. D. Sampson, L. Sheppard, T. Lumley, S. D. Adar, and J. Kaufman. Predicting Intra-Urban Variation in Air Pollution Concentrations with Complex Spatio-Temporal Dependencies. *Environmetrics*, 21(6):606–631, Sept. 2009.

[72] S. D. Team. *Stan UserâĂŹs Guide*. 2018.

[73] C. Tebaldi, R. L. Smith, D. Nychka, and L. O. Mearns. Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles. *Journal of Climate*, 18(10):1524–1540, May 2005.

[74] J. Tollefson. Air pollution science under siege at US environment agency. *Nature*, 568:15–16, Mar. 2019.

[75] R. Tuo and C. F. J. Wu. Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43(6):2331–2352, Dec. 2015.

[76] C. A. S. A. C. U.S. Environmental Protection Agency. CASAC Review of the EPAâĂŹs Integrated Science Assessment for Particulate Matter. Technical Report EPA-CASAC-19-002, U.S. Environmental Protection Agency, Washington, D.C., Apr. 2019.

[77] A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *arXiv:0805.3252 [math, stat]*, pages 200–222, 2008. arXiv: 0805.3252.

[78] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian Estimation Using a Gaussian Random Field with Inverse Gamma Bandwidth. *The Annals of Statistics*, 37(5B):2655–2675, 2009.

[79] A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2–95–2119, 2011.

[80] A. van Donkelaar, R. V. Martin, R. J. D. Spurr, and R. T. Burnett. High-Resolution Satellite-Derived PM2.5 from Optimal Estimation and Geographically Weighted Regression over North America. *Environmental Science & Technology*, 49(17):10482–10491, Sept. 2015.

[81] J. Wang and G. Song. A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction. *Neurocomputing*, 314:198–206, Nov. 2018.

[82] M. Wang, R. Beelen, X. Basagana, T. Becker, G. Cesaroni, K. de Hoogh, A. Dedele, C. Declercq, K. Dimakopoulou, M. Eeftens, F. Forastiere, C. Galassi, R. GraÅ¿uleviÄ■ienÄŮ, B. Hoffmann, J. Heinrich, M. Iakovides, N. KÃijnzli, M. Korek, S. Lindley, A. MÃűlter, G. Mosler, C. Madsen, M. Nieuwenhuijsen, H. Phuleria, X. Pedeli, O. Raaschou-Nielsen, A. Ranzi, E. Stephanou, D. Sugiri, M. Stempfelet, M.-Y. Tsai, T. Lanki, O. Udvardy, M. J. VarrÃş, K. Wolf, G. Weinmayr, T. Yli-Tuomi, G. Hoek, and B. Brunekreef. Evaluation of land use regression models for NO2 and particulate matter in 20 European study areas. *Environmental Science & Technology*, 47(9):4357–4364, May 2013.

[83] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[84] X. Wu, D. Braun, J. Schwartz, M. Kioumourtzoglou, and F. Dominici. Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. *Science advances*, 6(29):eaba5692, 2020.

[85] Q. Xiao, H. H. Chang, G. Geng, and Y. Liu. An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. *Environmental Science & Technology*, Oct. 2018.

[86] J. D. Yanosky, C. J. Paciorek, F. Laden, J. E. Hart, R. C. Puett, D. Liao, and H. H. Suh. Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. *Environmental Health*, 13(1):63, Aug. 2014.

[87] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, 13(3):917–1003, Sept. 2018.

A Expression for BNE's predictive mean

(**jzl:** To update) Recall the "Darth Vadar rule" [57]:

$$E(s(y)|\mathbf{x}) = \int_{y\in\mathscr{Y}} \frac{\partial}{\partial y} s(y) * \big[I(y>0) - F(y|\mathbf{x})\big] dy \tag{16}$$

Also recall that BNE's model CDF is $F(y|\mathbf{x}, G, \Phi) = G\Big[F_S(y|\mathbf{x}, \Phi)\Big]$. Therefore we can expresse the predictive mean $E(y|\mathbf{x}, G, \Phi)$ for full BNE in terms of its CDF:

$$\begin{aligned}
E(y|\mathbf{x}, G, \Phi) &= \int_{y\in\mathscr{Y}} I(y>0) - F(y|\mathbf{x}, G, \Phi) dy \\
&= \int_{y\in\mathscr{Y}} I(y>0) - G\Big[F_S(y|\mathbf{x}, \Phi)\Big] dy \\
&= \int_{y\in\mathscr{Y}} \Big[I(y>0) - F_S(y|\mathbf{x}, \Phi)\Big] + \Big[F_S(y|\mathbf{x}, \Phi) - G\big[F_S(y|\mathbf{x}, \Phi)\big]\Big] dy \\
&= \underline{\int_{y\in\mathscr{Y}} \Big[I(y>0) - F_S(y|\mathbf{x}, \Phi)\Big] dy} + \int_{y\in\mathscr{Y}} \Big[F_S(y|\mathbf{x}, \Phi) - G\big[F_S(y|\mathbf{x}, \Phi)\big]\Big] dy
\end{aligned}$$

Notice in the last line of above expression, the first integral (underlined) is the predictive mean with respect to the additive ensemble model $Y = \sum_{k=1}^{K} f_k(\mathbf{x})\mu_k + \delta(\mathbf{x}) + \varepsilon$. Therefore:

$$\begin{aligned}
E(y|\mathbf{x}, G, \Phi) &= \sum_{k=1}^{K} f_k(\mathbf{x})\mu_k + \underbrace{\delta(\mathbf{x})}_{D_\delta(y|\mathbf{x})} + \underbrace{\int_{y\in\mathscr{Y}} \Big[F_S(y|\mathbf{x}, \Phi) - G\big[F_S(y|\mathbf{x}, \Phi)\big]\Big] dy}_{D_G(y|\mathbf{x})} \\
&= \sum_{k=1}^{K} f_k(\mathbf{x})\mu_k + D_\delta(y|\mathbf{x}) + D_G(y|\mathbf{x})
\end{aligned} \tag{17}$$

B Scalable Gaussian Process Computation via Random-feature Parameterization

As discussed in Section 4, the computation of the Gaussian process posterior often involves inverting a $n \times n$ kernel matrix over the entire dataset, leading to a $O(n^3)$ computational complexity and making it infeasible for large-scale spatio-temporal problems. To this end, [63] has proposed a random-feature based parameterization of the Gaussian process prior that effectively reduced GP to a Bayesian linear model, hence allowing for scalable $O(n)$ computation when combined with a gradient-based MCMC procedure such as NUTS [32] or the Stochastic-gradient Laungevin Dynamics (SGLD) [83].

Specifically, consider a Gaussian process prior over data $\{\mathbf{x}_i\}_{i=1}^n$: $\mathbf{f} \sim MVN(0, \mathbf{K}_{n\times n})$, where $\mathbf{K}_{i,j} = k(\mathbf{x}_i - \mathbf{x}_j)$ is a $n \times n$ matrix for a shift invariant (e.g., the RBF kernel $k_{RBF}(\mathbf{x}_i - \mathbf{x}_j) = exp(-||\mathbf{x}_i - \mathbf{x}_j||_2^2/l)$). The random-feature method shows that a GP prior can be equivalently written in a linear model form:

$$\mathbf{f} = \mathbf{H}_{n\times D}\beta_{D\times 1}, \quad \beta_{D\times 1} \sim MVN(\mathbf{0}, \mathbf{I}_{D\times D}), \tag{18}$$

where $\mathbf{H}_{n\times D} = \frac{1}{\sqrt{D}}[h(\mathbf{x}_1)_{D\times 1}, \ldots, h(\mathbf{x}_n)_{D\times 1}]^\top$ is the *random-feature* transformation of the data $\mathbf{x}$. Note that (18) can be equivalent written as:

$$f(\mathbf{x}) \sim MVN(\mathbf{0}, \hat{\mathbf{K}} = \mathbf{H}\mathbf{H}^\top),$$

where

$$\hat{\mathbf{K}}_{i,j} = \hat{k}(\mathbf{x}_i - \mathbf{x}_j) = \frac{1}{D} h(\mathbf{x}_j)^\top h(\mathbf{x}_j) = \frac{1}{D}\sum_{p=1}^{D} h_p(\mathbf{x}_i) h_p(\mathbf{x}_j) = \hat{E}_p(h_p(\mathbf{x}_i) h_p(\mathbf{x}_j)). \tag{19}$$

is the approximated kernel function.

The transformation function $h : \mathbb{R}^d \to \mathbb{R}^D$ takes the below form:

$$h(\mathbf{x}) = \sqrt{2} * cos(\mathbf{W}_{D\times d}\mathbf{x}_{d\times 1} + \mathbf{b}_{D\times 1}) \tag{20}$$

where elements of $\mathbf{b}$ are drawn i.i.d. from $Unif(0, 2\pi)$. The elements of $\mathbf{W}_{D\times d}$ are drawn from $f(w)$, which is the Fourier transform of the shift-invariant kernel $k(\delta)$ for $\delta = \mathbf{x} - \mathbf{y}$. As a result, the random-feature formulation of (20) leads to a strong approximation guarantee of the GP kernel $k(\mathbf{x}_i - \mathbf{x}_j)$. That is, by the classic Bohner theorem, any shift-invariant kernel k can be expressed as the Fourier transformation [64]:

$$k(\mathbf{x}_i - \mathbf{x}_j) = Re\Big(\int_{w\in\mathbb{R}^d} exp(-zw^\top(\mathbf{x}_i - \mathbf{x}_j)) f(w) dw\Big) = E_w(h_w(\mathbf{x}_i)^\top h_w(\mathbf{x}_j))$$

where z is the imaginary unit and $h_w(\mathbf{x}) = cos(w^\top \mathbf{x} + b)$ for $w \sim f(w)$ and $b \sim Unif(0, 2\pi)$. This shows that (19) is indeed a valid, Monte-carlo-based approximation of the true kernel function k with a fast exponential rate of convergence in terms of the number of random features D [63, 50].

Consequently, to approximate a Gaussian process model with shift-invariant kernel $k(\delta)$, one only need to derive its Fourier density $f(w)$ and construct a D-dimensional linear model following (18) and (20). Many classic kernels k correspond to well-known distributions f. For example, the radial basis function (RBF) kernel corresponds to Gaussian distribution, while the Matérn kernel corresponds to Student's t-distribution [9]. Due to its simplicity and strong empirical performance, the random feature method has seen widespread adoption in both the applied and theoretical machine learning literature for building large-scale, high-dimensional probabilistic models [50, 12, 51, 24, 55, 54].

Coming back to the BNE model, recall the full BNE model likelihood presented in Section 4:

$$\begin{aligned}
&y|\mathbf{x} \sim N(\sum_k \omega_k(\mathbf{x}) f(\mathbf{x}) + \delta(\mathbf{x}) + m(\mathbf{x}), \sigma^2 * s^2(\mathbf{x})) \\
&\omega = \frac{exp(w_k)}{\sum_k exp(w_k)}, \quad w_k \sim GP(0, k_w) \\
&\delta \sim GP(0, k_\delta), \quad log\, s \sim GP(0, \sigma_s^2 * k_s) \\
&m \sim Exp(\lambda), \quad log\, \lambda \sim GP(\mu_\lambda, \sigma_\lambda^2 * k_\lambda).
\end{aligned}$$

Using the random-feature parametrization, the GP priors in the original BNE model are reduced to:

$$\begin{aligned} w_k(\mathbf{x}) &= h(\mathbf{x})^\top \beta_k, & \beta_{k,D\times 1} &\sim MVN(0,\mathbf{I}) \\ \delta(\mathbf{x}) &= h(\mathbf{x})^\top \beta_\delta, & \beta_{\delta,D\times 1} &\sim MVN(0,\mathbf{I}) \\ log\, s(\mathbf{x}) &= h(\mathbf{x})^\top \beta_s, & \beta_{s,D\times 1} &\sim MVN(0,\sigma_s^2\mathbf{I}) \\ log\, \lambda(\mathbf{x}) &= h(\mathbf{x})^\top \beta_\lambda + \mu_\lambda, & \beta_{\lambda,D\times 1} &\sim MVN(0,\sigma_\lambda^2\mathbf{I}), \end{aligned}$$

As a result, we only need to perform posterior inference with respect to K+3 Bayesian linear models $(\{\beta_k\}_{k=1}^K, \beta_\delta, \beta_s, \beta_\lambda)$ with fixed dimensions D, reducing the computation complexity from $O(n^3)$ to $O(n)$. Furthermore, due to the low-dimensionality of spatio-temporal modeling (i.e., d=3), a small amount of random features (e.g., $D = 128$) is already sufficient of obtaining high-quality approximation to the GP posterior.

C Theoretical Properties

C.1 Prediction

Recall that the goal of a BNE is to produce both calibrated predictive uncertainty and accurate prediction. Here, we show in Theorem 1 that *calibration preserves accuracy*. That is, when compared to an uncalibrated model (i.e. the model estimated only using Step 1 and 2), the calibrated model's excess predictive error in terms of the MSE is bounded and asymptotically approaches zero. Intuitively, this result implies that for reasonably large sample sizes, the calibrated model cannot have worse predictive accuracy than the uncalibrated model.

Theorem 1 (Calibration Preserves Accuracy, MSE Loss)
Denote $F_\mu \in \mathscr{F}$ the uncalibrated predictive CDF, and $F = G\big[F_\mu\big] \in \mathscr{F}$ the calibrated predictive CDF. Given observations $Y = \{y_i\}_{i=1}^N$ generated from $F^*(y|\mathbf{x}_i)'s$, we denote $\mathbb{L}(Y,F) = \frac{1}{N}\sum_{i=1}^n (y_i - E(y|\mathbf{x}_i))^2$ the MSE loss, and denote $\mathbb{C}(Y,F)$ the calibration loss:

$$\mathbb{C}(Y,\hat{F}) = \int_{y\in\mathscr{Y}} \frac{1}{N}\sum_{i=1}^n \Big|\mathbb{I}(y_i < y|\mathbf{x}_i) - F(y|\mathbf{x}_i)\Big| dy.$$

i.e. the empirical L_1 loss between model and empirical CDFs.

Assume that F is a consistent estimator for F^*, i.e. $\underset{n\to\infty}{limsup}\, \mathbb{C}(Y,\hat{F}) \leq \varepsilon_n$ for a sequence $\varepsilon_n \to 0$. The excess error of the calibrated CDF F with respect to the uncalibrated CDF F_μ in terms of the MSE loss $\mathbb{L}$ is bounded by the asymptotically vanishing term ε_n up to a constant C:

$$\underset{n\to\infty}{limsup}\Big[\mathbb{L}(Y,\hat{F}) - \mathbb{L}(Y,\hat{F}_0)\Big] \leq C\varepsilon_n \tag{21}$$

Proof. See Section C.2 □

Specifically, the expression in (21) suggests that a calibration model with faster convergence rate (i.e. smaller ε_n) leads to greater improvement in the prediction performance (in the sense that the excess error is small). To this end, we notice that by using an Matérn $\frac{3}{2}$ kernel with the length-scale parameter selected adaptively from data, ε_n may reach the minimax-optimal rate up to a logarithmic factor [78], leading to improved prediction performance in finite samples.

C.2 Proof for Theorem 1

Theorem 1 is a special case of Theorem 2 when the accuracy measure is MSE. We present Theorem 2 below.

Theorem 2 (Calibration Preserves Accuracy, General Loss)
Denote $F_0 \in \mathscr{F}$ the uncalibrated predictive CDF, and $F = G \circ F_0 \in \mathscr{F}$ the calibrated predictive CDF. Given observations $\{y_i, \mathbf{x}_i\}_{i=1}^N$, we denote:

1. $\mathbb{L}(Y,F) = \frac{1}{N}\sum_{i=1}^N L(y_i, F(Y|\mathbf{X}=\mathbf{x}_i))$ *with respect to a accuracy measure function $L : \mathscr{Y} \times \mathscr{F} \to \mathbb{R}$ that is bounded and proper, i.e.:*

   (a) ***Bounded**: There exists positive constant $B < \infty$ such that $L(y,F) < B \quad \forall y \in \mathscr{Y}, F \in \mathscr{F}$. Furthermore, for $F, F' \in \mathscr{F}$ such that $\int_{\mathscr{Y}} \Big|F(t) - F'(t)\Big| dt < \varepsilon$, we have:*

   $$\int_{\mathscr{Y}} \Big|L(y,F') - L(y,F)\Big| dy < B\varepsilon.$$

   (b) ***Proper**: $\mathbb{L}(Y,F)$ is minimized by the true CDF $\mathbb{F}$, i.e. $\mathbb{L}(Y,\mathbb{F}) \leq \mathbb{L}(Y,F) \quad \forall F \in \mathscr{F}$.*

2. $\mathbb{C}(Y,\hat{F})$ *the empirical calibration loss defined as:*

   $$\mathbb{C}(Y,\hat{F}) = \int_{\mathscr{Y}} \frac{1}{N}\sum_{i=1}^n \Big|\mathbb{I}(y_i < t|\mathbf{x}_i) - F(t|\mathbf{x}_i)\Big| dt$$

Furthermore, we assume that $\hat{F}$ is ε_n-calibrated [1], i.e. $\underset{n\to\infty}{limsup}\, \mathbb{C}(Y,\hat{F}) \leq \varepsilon_n$ for some $\varepsilon_n > 0$.

Then the excess error of the calibrated CDF $\hat{F}$ with respect to the uncalibrated CDF $\hat{F}_0$ in terms of $\mathbb{L}$ is bounded by $\mathbb{C}$, i.e.

$$\mathbb{L}(Y,\hat{F}) - \mathbb{L}(Y,\hat{F}_0) \leq 2B * \mathbb{C}(Y,\hat{F}) + (B+1) * \varepsilon_n$$

Morever, since $\hat{F}$ is ε_n-calibrated:

$$\underset{n\to\infty}{limsup}\Big[\mathbb{L}(Y,\hat{F}) - \mathbb{L}(Y,\hat{F}_0)\Big] \leq (3B+1) * \varepsilon_n \tag{22}$$

Our proof technique is an adaptation of the game-theoretic analysis of internal regret from the setting of online sequential classification [7, 45], and does not assume independence between the observations.

Proof. For $Y, \mathbf{X} \sim \mathbb{F}(Y, \mathbf{X})$, denote $l(Y,F)(t) = L(Y, F(t|\mathbf{X}))$ the accuracy measure that involves the observation Y and the corresponding predictive CDF $F(Y|\mathbf{X})$ evaluated at $t \in \mathscr{Y}$. For an observed data pair $\{y_i, \mathbf{x}_i\}$, denote the corresponding empirical version of l as $l_i(Y,F)(t) = L(y_i, F(t|\mathbf{x}_i))$. Also denote $F_i(t) = F(t|\mathbf{X} = \mathbf{x}_i)$.

We notice below facts:

1. $l(Y,F)(t)$ is bounded by B, therefore for any predictive CDFs $F, F' \in \mathscr{F}$, we always have:
$$l(Y,F)(t) - l_i(Y,F')(t) \leq B$$

2. Since $l(Y,F)(t)$ is proper in the sense that this loss is minimized by the true model parameters, then for $Y \sim \mathbb{F}$:
$$\mathbb{F}(t) \in \underset{F \in \mathscr{F}}{argmin}\, E_{Y \sim \mathbb{F}}\Big(l(Y,F)(t)\Big)$$

3. $\mathbb{I}\Big(\hat{F}_i(t) = p_i\Big) = \mathbb{I}\Big(\hat{F}_i(t) = p_i, y_i < t\Big) + \mathbb{I}\Big(\hat{F}_i(t) = p_i, y_i > t\Big)$.
Furthermore,
$$\mathbb{I}\Big(\hat{F}_i(t) = p_i, y_i < t\Big) = \mathbb{I}\Big(\hat{F}_i(t) = p_i\Big) * \Big(\mathbb{I}\big(y_i < t|\hat{F}_i(t) = p_i\big) - p_i\Big) + \mathbb{I}\Big(\hat{F}_i(t) = p_i\Big) * p_i$$
$$\mathbb{I}\Big(\hat{F}_i(t) = p_i, y_i > t\Big) = \mathbb{I}\Big(\hat{F}_i(t) = p_i\Big) * \Big(p_i - \mathbb{I}\big(y_i < t|\hat{F}_i(t) = p_i\big)\Big) + \mathbb{I}\Big(\hat{F}_i(t) = p_i\Big) * \Big(1 - p_i\Big)$$
which implies:
$$\mathbb{I}\Big(\hat{F}_i(t) = p_i\Big) \leq 2 * \mathbb{I}\Big(\hat{F}_i(t) = p_i\Big) * \Big|\mathbb{I}\big(y_i < t|\hat{F}_i(t) = p_i\big) - p\Big| + \mathbb{I}\Big(\hat{F}_i(t) = p_i\Big) * p_i(t)$$
where $p_i(t) = p_i$ if $y_i < t$ and $p_i(t) = 1 - p_i$ otherwise.

Combining above facts, we have:
$$\begin{aligned}
\mathbb{L}(Y,\hat{F})(t) - \mathbb{L}(Y,\hat{F}_0)(t) &= \frac{1}{N}\sum_{i=1}^{N}\Big(l_i(Y,\hat{F})(t) - l_i(Y,\hat{F}_0)(t)\Big) * \mathbb{I}(\hat{F}_i(t) = p_i) \\
&\leq 2\frac{1}{N}\sum_{i=1}^{N}\underline{\Big(l_i(Y,\hat{F})(t) - l_i(Y,\hat{F}_0)(t)\Big)} * \Big|\mathbb{I}\big(y_i < t|\hat{F}_i(t) = p_i\big) - p_i\Big| * \mathbb{I}(\hat{F}_i(t) = p_i) + \\
&\quad \frac{1}{N}\sum_{i=1}^{N}\Big(l_i(Y,\hat{F})(t) - l_i(Y,\hat{F}_0)(t)\Big) * p_i(t) * \mathbb{I}(\hat{F}_i(t) = p_i) \\
&\leq 2B * \left\{\frac{1}{N}\sum_{i=1}^{N}\Big|\mathbb{I}\big(y_i < t|\hat{F}_i(t) = p_i\big) - p_i\Big| * \mathbb{I}(\hat{F}_i(t) = p_i)\right\} + \\
&\quad \left\{\frac{1}{N}\sum_{i=1}^{N}\Big(l_i(Y,\hat{F})(t) - l_i(Y,\hat{F}_0)(t)\Big) * p_i(t) * \mathbb{I}(\hat{F}_i(t) = p_i)\right\} \qquad (23)
\end{aligned}$$

Here the first inequality uses Fact 3. The second inequality uses Fact 1 on the underlined component.

We now consider the two expressions in curly brackets in (23). Notice the first expression corresponds to an "unintegrated" version of calibration loss:

$$\frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{I}\big(y_i < t|\hat{F}_i(t) = p_i\big) - p_i\right| * \mathbb{I}(\hat{F}_i(t) = p_i) = \frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{I}\big(y_i < t|\mathbf{x}_i\big) - \hat{F}_i(t)\right| = \mathbb{C}(Y,\hat{F})(t), \tag{24}$$

and the second expression can be written as:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\Big(l_i(Y,\hat{F})(t) - l_i(Y,\hat{F}_0)(t)\Big) * p_i(t) * \mathbb{I}(\hat{F}_i(t) = p_i) \\
&= \frac{1}{N}\left[\sum_{y_i<t}\Big(l_i(Y,\hat{F})(t) - l_i(Y,\hat{F}_0)(t)\Big) * \hat{F}_i(t) + \sum_{y_i\geq t}\Big(l_i(Y,\hat{F})(t) - l_i(Y,\hat{F}_0)(t)\Big) * (1-\hat{F}_i(t))\right] \\
&\leq \frac{1}{N}\left[\sum_{y_i<t}\Big(l_i(Y,\hat{F})(t) - l_i(Y,\hat{F}_0)(t)\Big) * \mathbb{I}\big(y_i < t|\mathbf{x}_i\big) + \sum_{y_i\geq t}\Big(l_i(Y,\hat{F})(t) - l_i(Y,\hat{F}_0)(t)\Big) * \mathbb{I}\big(y_i > t|\mathbf{x}_i\big)\right] + \varepsilon_n \\
&= \mathbb{E}\Big(l(Y,\hat{F})(t) - l(Y,\hat{F}_0)(t)\Big) + \varepsilon_n \\
&\leq \mathbb{E}\Big(l(Y,\mathbb{F})(t) - l(Y,\hat{F}_0)(t)\Big) + (B+1)\varepsilon_n \\
&\leq (B+1)\varepsilon_n
\end{aligned} \tag{25}$$

where the first inequality follows since $\hat{F}$ is ε_n-calibrated, the second inequality follows since l is a bounded loss, and the last inequality follows since l is a proper loss.

Now replace the two expressions within brackets in (23) with (24) and (25), we have:

$$\mathbb{L}(Y,\hat{F})(t) - \mathbb{L}(Y,\hat{F}_0)(t) \leq 2B * \mathbb{C}(Y,\hat{F})(t) + (B+1)\varepsilon_n$$

Intergrating both sides of above equation with respect to $t \in \mathscr{Y}$, we have:

$$\mathbb{L}(Y,\hat{F}) - \mathbb{L}(Y,\hat{F}_0) \leq 2B * \mathbb{C}(Y,\hat{F}) + (B+1)\varepsilon_n$$

□

Notice that Theorem 2 is applicable to most of the standard measures for predictive accuracy, e.g. root mean squared error (RMSE). This is because for suitably standardized $\{y_i\}_{i=1}^N$, RMSE is always bounded and proper [27].

Table 1: Summary of ensemble methods for simulation experiments in Section 5

Model Name	Label	Adaptive Ensemble	Bias Correction	Distribution Calibration
Bayesian Stacking	**Stacking**			
Generalized Additive Ensemble	**GAM**		Yes	
Adaptive Bayesian Model Averaging	**BMA**	Yes		
Bayesian Adaptive Ensemble	**BAE**	Yes	Yes	
Bayesian Nonparametric Ensemble	**BNE**	Yes	Yes	Yes

Table 2: Mean and Standard Deviation for RMSE and ECE in 1D time-series regression task.

Metric	Model	$n = 10$	$n = 25$	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
Prediction Metrics									
MSE	**Stacking**								
	GAM								
	BMA	0.4195	0.4159	0.3894	0.3548	0.3399	0.3246	0.3085	0.3109
	BAE	0.2562	**0.2354**	0.1917	0.1464	0.1086	0.0757	0.0657	0.0561
	BNE	**0.2492**	0.2439	**0.1816**	**0.1370**	**0.1029**	**0.0713**	**0.0606**	**0.0507**
Calibration Metrics									
NLL	**Stacking**								
	GAM								
	BMA	0.8706	0.8508	0.8474	0.8048	0.8031	0.7761	0.7751	0.7814
	BAE	0.7489	0.7251	0.7030	0.6557	0.6603	0.6335	0.6417	0.6410
	BNE	**0.5545**	**0.5103**	**0.4597**	**0.3853**	**0.2765**	**0.2386**	**0.2256**	**0.2238**
Coverage Probability Metrics									
ECE	**Stacking**								
	GAM								
	BMA	0.0064	0.0072	0.0070	0.0067	0.0075	0.0075	0.0075	0.0075
	BAE	0.0040	0.0046	0.0045	0.0042	0.0049	0.0051	0.0055	0.0052
	BNE	**0.0023**	**0.0023**	**0.0020**	**0.0017**	**0.0013**	**0.0014**	**0.0015**	**0.0014**

Table 3: Mean and Standard Deviation for RMSE and ECE in 2D spatial regression task.

Metric	Model	$n = 25$	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
Prediction Metrics								
MSE	**Stacking**							
	GAM							
	BMA	0.2087	0.2008	0.1887	0.1725	0.1607	0.1561	0.1569
	BAE	**0.1558**	**0.1480**	**0.1379**	**0.1250**	**0.1164**	**0.1137**	**0.1136**
	BNE	0.1566	0.1498	0.1397	0.1266	0.1189	0.1173	0.1169
Calibration Metrics								
NLL	**Stacking**							
	GAM							
	BMA	-2.1288	-2.1632	-2.1988	-2.2587	-2.2914	-2.3066	-2.2892
	BAE	-2.4281	-2.4555	-2.4836	-2.5292	-2.5582	-2.5720	-2.5694
	BNE	**-2.5229**	**-2.5466**	**-2.5772**	**-2.6235**	**-2.6643**	**-2.6799**	**-2.6841**
Coverage Probability Metrics								
ECE	**Stacking**							
	GAM							
	BMA	0.0132	0.0131	0.0125	0.0121	0.0115	0.0114	0.0117
	BAE	0.0056	0.0054	0.0051	0.0046	0.0043	0.0041	0.0042
	BNE	**0.0044**	**0.0042**	**0.0039**	**0.0036**	**0.0034**	**0.0033**	**0.0035**

Table 4: Mean and Standard Deviation for leave-one-out root mean squared error (RMSE) for annual $PM_{2.5}$ predictions.

Metric / Model	**Stacking**	**GAM**	**BMA**	**BAE**	**BNE**
RMSE	1.677 ± 0.124	1.544 ± 0.128	1.233 ± 0.127	1.077 ± 0.157	0.758 ± 0.088

Table 5: Full results for the Calibration and Coverage probability metrics on out-of-sample data from the training time window $(-3\pi, 3\pi)$ and the test time window $(-5\pi, 5\pi)$ for the time-series regression task. **IND**: "IN-Domain" result from the training time window. **ALL**: full result from the entire test time window.

Metric	Model	$n = 10$	$n = 25$	$n = 50$	$n = 100$	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
Calibration Metrics									
NLL (IND)	**Stacking**								
	GAM								
	BMA	0.5056	0.5009	0.4846	0.4334	0.4615	0.4511	0.4439	0.4477
	BAE	0.2733	0.2672	0.2224	0.1752	0.2092	0.2062	0.2204	0.2099
	BNE	0.1105	0.0519	-0.0529	-0.1869	-0.3569	-0.4275	-0.4452	-0.4668
NLL (ALL)	**Stacking**								
	GAM								
	BMA	0.8706	0.8508	0.8474	0.8048	0.8031	0.7761	0.7751	0.7814
	BAE	0.7489	0.7251	0.7030	0.6557	0.6603	0.6335	0.6417	0.6410
	BNE	0.5545	0.5103	0.4597	0.3853	0.2765	0.2386	0.2256	0.2238
Calibration (IND)	**Stacking**								
	GAM								
	BMA	0.5064	0.4087	0.3829	0.2923	0.2347	0.2078	0.1889	0.1913
	BAE	0.4282	0.3191	0.2504	0.1543	0.0862	0.0561	0.0465	0.0402
	BNE	0.2806	0.2677	0.2453	0.2004	0.1287	0.0873	0.0747	0.0621
Calibration (ALL)	**Stacking**								
	GAM								
	BMA	0.6148	0.5300	0.5153	0.4526	0.3906	0.3546	0.3321	0.3405
	BAE	0.6066	0.5118	0.4678	0.3958	0.3297	0.2862	0.2630	0.2687
	BNE	0.3552	0.3513	0.3487	0.3448	0.3207	0.3150	0.2979	0.3055
Sharpness (IND)	**Stacking**								
	GAM								
	BMA	-0.0009	0.0922	0.1017	0.1411	0.2268	0.2433	0.2550	0.2565
	BAE	-0.1549	-0.0519	-0.0279	0.0209	0.1230	0.1501	0.1740	0.1696
	BNE	-0.1701	-0.2157	-0.2982	-0.3873	-0.4856	-0.5149	-0.5198	-0.5289
Sharpness (ALL)	**Stacking**								
	GAM								
	BMA	0.2558	0.3208	0.3321	0.3522	0.4125	0.4216	0.4430	0.4408
	BAE	0.1423	0.2133	0.2353	0.2599	0.3305	0.3473	0.3787	0.3723
	BNE	0.1994	0.1590	0.1111	0.0405	-0.0442	-0.0764	-0.0723	-0.0817
Coverage Probability Metrics									
ECE (IND)	**Stacking**								
	GAM								
	BMA	0.0065	0.0072	0.0073	0.0073	0.0079	0.0082	0.0080	0.0083
	BAE	0.0037	0.0043	0.0046	0.0045	0.0053	0.0057	0.0059	0.0059
	BNE	0.0021	0.0021	0.0019	0.0015	0.0013	0.0014	0.0015	0.0016
ECE (ALL)	**Stacking**								
	GAM								
	BMA	0.0064	0.0072	0.0070	0.0067	0.0075	0.0075	0.0075	0.0075
	BAE	0.0040	0.0046	0.0045	0.0042	0.0049	0.0051	0.0055	0.0052
	BNE	0.0023	0.0023	0.0020	0.0017	0.0013	0.0014	0.0015	0.0014
95% CI Coverage Probability (IND)	**Stacking**								
	GAM								
	BMA	0.9129	0.9265	0.9303	0.9397	0.9478	0.9508	0.9532	0.9527
	BAE	0.8996	0.9158	0.9227	0.9339	0.9494	0.9564	0.9583	0.9583
	BNE	0.9257	0.9258	0.9305	0.9312	0.9471	0.9562	0.9592	0.9615
95% CI Coverage Probability (ALL)	**Stacking**								
	GAM								
	BMA	0.9178	0.9284	0.9322	0.9382	0.9461	0.9496	0.9523	0.9510
	BAE	0.9066	0.9184	0.9241	0.9321	0.9441	0.9503	0.9533	0.9521
	BNE	0.9460	0.9452	0.9475	0.9443	0.9511	0.9519	0.9544	0.9546

Table 6: Full results for the Calibration and Coverage probability metrics on out-of-sample data from the training region $(-\pi,\pi)^2$ and the test region $(-1.25\pi,1.25\pi)^2$ for the spatial regression task. **IND**: "IN-Domain" result from the training region. **ALL**: full result from the entire test region.

Metric	Model	$n=25$	$n=50$	$n=100$	$n=250$	$n=500$	$n=750$	$n=1000$
Calibration Metrics								
NLL (IND)	**Stacking**							
	GAM							
	BMA	-2.8318	-2.8464	-2.8708	-2.9216	-2.9458	-2.9679	-2.9731
	BAE	-3.0345	-3.0525	-3.0711	-3.1191	-3.1381	-3.1564	-3.1649
	BNE	-3.1267	-3.1407	-3.1648	-3.2149	-3.2457	-3.2569	-3.2695
NLL (ALL)	**Stacking**							
	GAM							
	BMA	-2.1288	-2.1632	-2.1988	-2.2587	-2.2914	-2.3066	-2.2892
	BAE	-2.4281	-2.4555	-2.4836	-2.5292	-2.5582	-2.5720	-2.5694
	BNE	-2.5229	-2.5466	-2.5772	-2.6235	-2.6643	-2.6799	-2.6841
Calibration (IND)	**Stacking**							
	GAM							
	BMA	0.7644	0.7578	0.7406	0.7063	0.6993	0.6872	0.6853
	BAE	0.6560	0.6462	0.6368	0.6080	0.6065	0.5966	0.5917
	BNE	0.5917	0.5864	0.5724	0.5445	0.5359	0.5355	0.5298
Calibration (ALL)	**Stacking**							
	GAM							
	BMA	1.0430	1.0210	1.0000	0.9682	0.9683	0.9692	0.9888
	BAE	0.8098	0.7952	0.7834	0.7687	0.7740	0.7754	0.7807
	BNE	0.6964	0.6854	0.6698	0.6530	0.6448	0.6450	0.6444
Sharpness (IND)	**Stacking**							
	GAM							
	BMA	-3.5962	-3.6042	-3.6115	-3.6279	-3.6450	-3.6551	-3.6584
	BAE	-3.6905	-3.6988	-3.7079	-3.7271	-3.7446	-3.7529	-3.7566
	BNE	-3.7184	-3.7271	-3.7372	-3.7594	-3.7816	-3.7925	-3.7993
Sharpness (ALL)	**Stacking**							
	GAM							
	BMA	-3.1719	-3.1842	-3.1988	-3.2269	-3.2597	-3.2758	-3.2780
	BAE	-3.2379	-3.2507	-3.2671	-3.2980	-3.3323	-3.3474	-3.3501
	BNE	-3.2192	-3.2319	-3.2470	-3.2765	-3.3091	-3.3249	-3.3286
Coverage Probability Metrics								
ECE (IND)	**Stacking**							
	GAM							
	BMA	0.0056	0.0058	0.0059	0.0060	0.0061	0.0061	0.0062
	BAE	0.0043	0.0043	0.0043	0.0045	0.0046	0.0047	0.0049
	BNE	0.0048	0.0048	0.0049	0.0051	0.0051	0.0052	0.0054
ECE (ALL)	**Stacking**							
	GAM							
	BMA	0.0132	0.0131	0.0125	0.0121	0.0115	0.0114	0.0117
	BAE	0.0056	0.0054	0.0051	0.0046	0.0043	0.0041	0.0042
	BNE	0.0044	0.0042	0.0039	0.0036	0.0034	0.0033	0.0035
95% CI Coverage Probability (IND)	**Stacking**							
	GAM							
	BMA	0.9400	0.9390	0.9429	0.9445	0.9464	0.9443	0.9462
	BAE	0.9449	0.9454	0.9473	0.9507	0.9520	0.9517	0.9519
	BNE	0.9570	0.9577	0.9579	0.9616	0.9641	0.9618	0.9629
95% CI Coverage Probability (ALL)	**Stacking**							
	GAM							
	BMA	0.8486	0.8491	0.8536	0.8561	0.8586	0.8568	0.8574
	BAE	0.8762	0.8777	0.8792	0.8827	0.8846	0.8848	0.8848
	BNE	0.9298	0.9308	0.9325	0.9336	0.9360	0.9340	0.9336

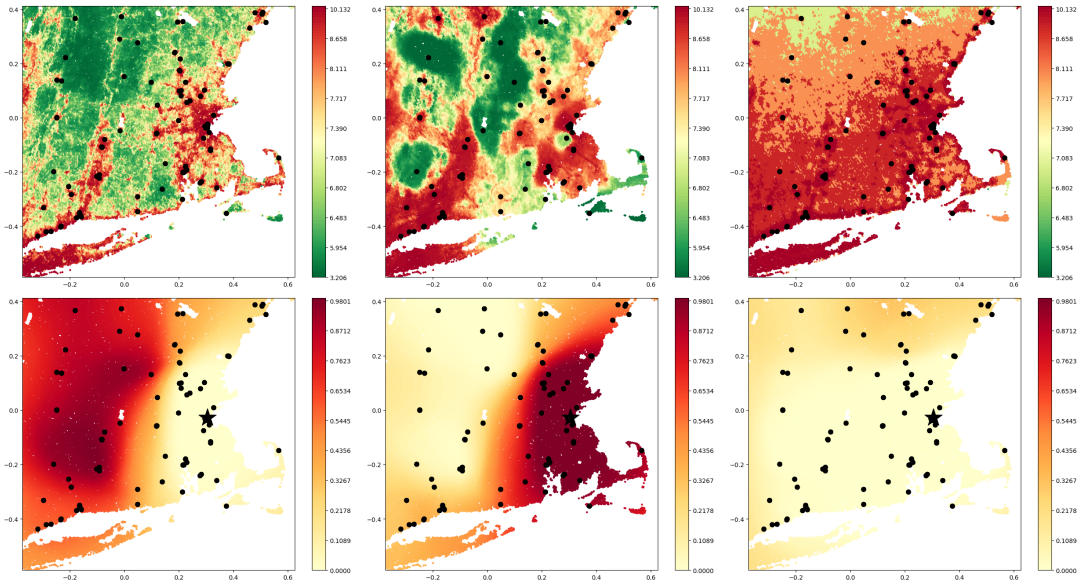


Figure 1: Annual average $PM_{2.5}$ predictions (Top) and posterior mean of ensemble weights (Bottom) assigned to each air pollution models in Eastern Massachusetts in 2011. **Left** Kloog et al. (2014) [42]; **Middle** Di et al. (2017) [18]; **Right** van Donkelaar et al. (2015) [80]; **Black Dots**: Monitor Locations; **Black Star**: Boston, MA.

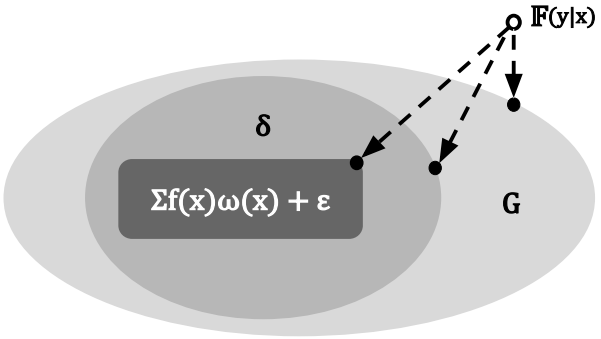


Figure 2: Illustration of the BNE's model space. **Dark Gray**: $\mathscr{M}_{\omega}$; **Medium Gray**: $\mathscr{M}_{\omega,\delta}$; **Light Gray**: $\mathscr{M}_{\omega,\delta,G}$; **Empty Dot**: Empirical Distribution $\mathbb{F}(y|\mathbf{x})$; **Black Dots**: Projection of $\mathbb{F}(y|\mathbf{x})$ to Model Spaces.

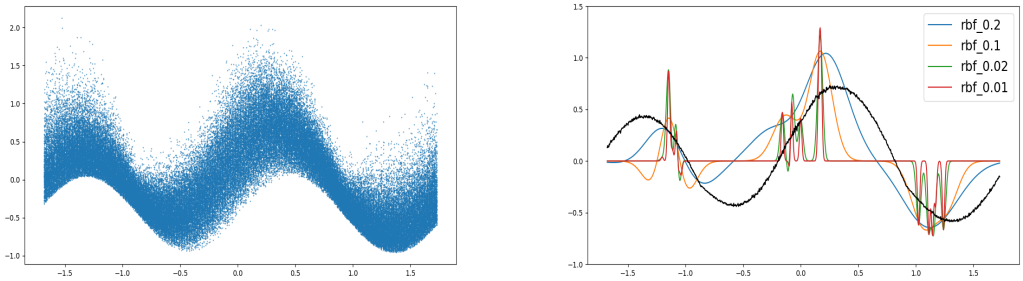


Figure 3: Data generation function (left) and the deterministic predictions from base models (right) in the time-series experiment. **Black Line**: data-generation function. **Colored Line**: Base model predictions.

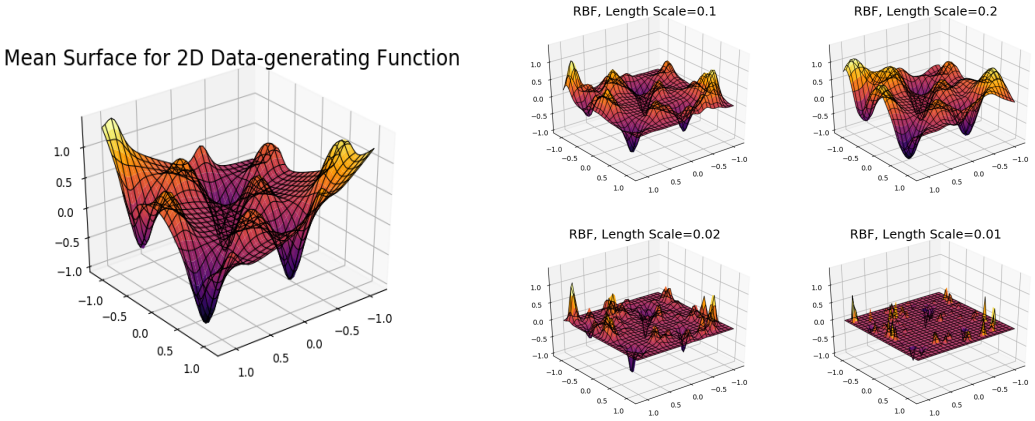


Figure 4: Samples from the data-generating distribution (left) and the deterministic predictions (right) from base models in the spatial experiment.

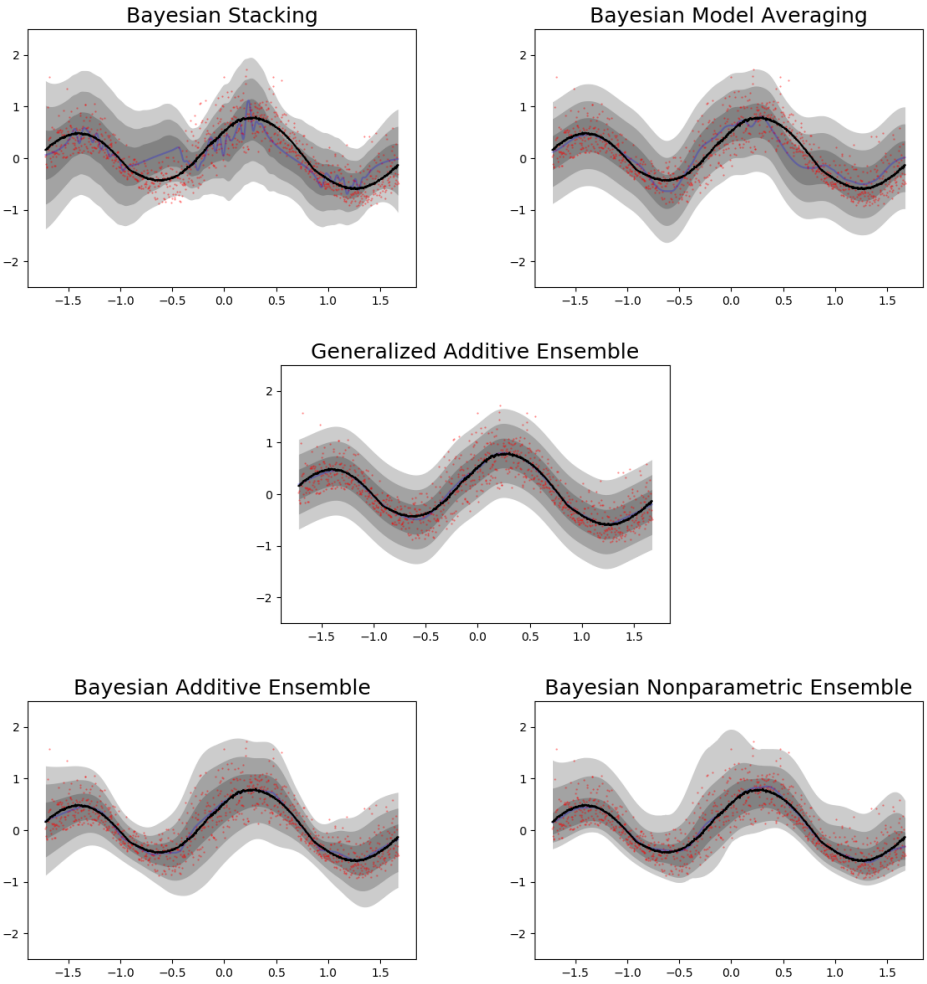
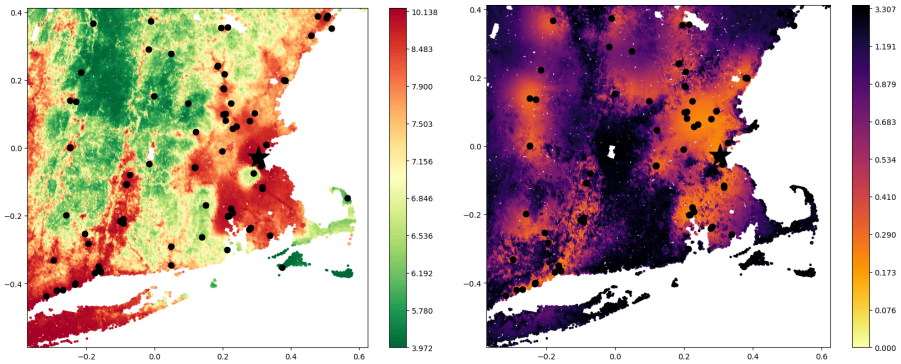


Figure 5: Comparison in prediction and uncertainty quantification of different ensemble methods for the Weibull-distributed time-series data data. **Red Dots**: Observations in the testing data. **Black Line**:Mean of data generating distribution. **Blue Line**: posterior predictive mean. **Grey Shade**: $[0.3\%, 5\%, 32\%, 68\%, 95\%, 99.7\%]$ quantiles of the model's posterior predictive distribution.



(a) Posterior Mean

(b) Overall Uncertainty

(c) Uncertainty, Model Selection

(d) Uncertainty, Prediction

Figure 6: Posterior predictive mean, predictive uncertainty (i.e. variance) and its decomposition in the BNE model. **Black Dots**: Location of air pollution monitors; **Black Star**: Boston, MA.