Adaptive and Calibrated Ensemble Learning with Dependent Tail-free Process

Jeremiah Zhe Liu$^1$  John Paisley$^2$  Marianthi-Anna Kioumourtzoglou$^2$  Brent A. Coull$^1$
$^1$Harvard University, Cambridge, MA, 02139
$^2$Columbia University, New York, NY, 10027
zh112@mail.harvard.edu, jpaisley@columbia.edu
mk3961@cumc.columbia.edu, bcoull@hsph.harvard.edu

1 Introduction

Ensemble learning is a mainstay in many modern machine-learning systems interacting with real-world data [14]. Conventional ensemble algorithms assign to base models a set of deterministic, constant model weights, not fully accounting for variability in the base model’s ability to capture different aspects of the data-generation mechanism, nor providing uncertainty estimates for the ensemble prediction. The motivating application for this work arises from the field of air pollution exposure assessment. To improve exposure assessment and minimize exposure measurement error, many different research groups are currently building highly resolved spatio-temporal prediction models. These different models have different inputs (from satellite remote sensing to chemical transport models and land use variables) and employ different algorithms (from linear mixed models, generalized additive models to neural networks). Although the aim of each of these models is usually to maximize the global predictive accuracy, in practice this accuracy varies both in space and in time (see, for example, Figure D.6). In this case, an ensemble algorithm assigning deterministic, spatially constant ensemble weights is neither accurate in prediction or informative in its prediction’s reliability. Therefore for an ensemble method to be accurate and reliable in such applications, it is crucial for the method to exhibit adaptivity, i.e. combine the base model predictions differently according their predictive performance over space and time, and also to provide a calibrated estimate of uncertainty, in the sense that the model’s predictive uncertainty faithfully reflects its actual likelihood of being correct. Recently, several ensemble approaches designed to improve spatio-temporal air pollution predictions have been developed, including hierarchical models with sophisticated covariance structure[18][12], bootstrap aggregation of regression trees or neural networks [13][9][21], and stacked generalization of multiple black-box algorithms[22]. However, to our knowledge, no method to date has integrated spatio-temporal weighing of the base models, i.e. has assign larger weights at each space and time point to the base model with the highest accuracy. Importantly, no method has provided comprehensive intra- and inter-model characterization of the spatio-temporal uncertainty in the predictions.

To address the above gaps in methodology, in this work we present a feature-adaptive, probabilistic approach to ensemble learning using a dependent tail-free process as the ensemble weight prior. Specifically, we model the ensemble weight as a random measure $\mu : \mathcal{F} \times \mathcal{X} \rightarrow [0, 1]$ that depends on the input feature $x$ (e.g. the spatial location), where $\mathcal{F}$ is the space of base models and $\mathcal{X}$ is the input feature space. Therefore, as $x \in \mathcal{X}$ changes, $\mu(\hat{f}, x)$ distributes the weights differently among the base models, while at the same time quantifying uncertainty in weight assignment due to the fact that we take these weights as random measures. The resulting method incorporates the prior information about the base models, and provides interpretable uncertainty quantification that can be decomposed into uncertainty due to model selection and that due to ensemble prediction.

To scale computation to the real-world sized problems without sacrificing the model’s ability to accurately quantify uncertainty, in Section 3, we develop a novel variational inference algorithm that orients the variational family toward both approximating the model posterior, and also toward producing a calibrated predictive distribution, in the sense that the predictive distribution should be
consistent with the empirical distribution of the data\cite{7}. We achieve this by defining the objective function to be a composition of two distance measures: a Kullback-Leibler (KL) distance that measures the variational family’s quality in approximating the model posterior, and an additional “calibration” distance that measures the variational predictive distribution’s consistency with the empirical distribution of the data. The specific calibration distance we consider is the Cramer-von Mises (CVM) distance \cite{1}, i.e. the \( L_2 \) distance between the model’s predictive cumulative distribution function (CDF) \( \hat{P}(y_i < t) \) and the data’s empirical CDF: \( \text{CvM}(\hat{P}, y_i) = \int_0^t [\hat{P}(y_i < t) - P(y_i < t)]^2 dt. \) As distance measures between probability distributions, both KL and CVM induce quality measures for model’s uncertainty quantification known as proper scoring rules \cite{3}. Compared to the KL distance, the CVM distance comprehensively access the entire predictive function’s quality in approximating the empirical distribution. Consequently, the Continuously Ranked Probability Score (i.e. the scoring rule induced by the CVM distance) is commonly preferred as a more robust alternative to logarithm score (i.e. the scoring rule induced by the KL)\cite{17, 7}. In Section 4 we empirically investigate the behavior of our method and compare its performance with that of the traditional ensemble methods on a nonlinear function regression benchmark. We conclude with a real-world application of spatial integration of \( PM_{2.5} \) pollution prediction models in the greater Boston area.

2 Model

Given observations \( \{x_i, y_i\}_{i=1}^N \) and a set of base models \( \mathcal{F} = \{\hat{f}_k\}_{k=1}^K \), we assume there exists an ensemble function \( f \) such that:

\[
y_i | f, x_i \sim N(f(x_i), \sigma^2), \quad f(x_i) = \sum_{k=1}^K \hat{f}_k(x_i) \mu(\hat{f}_k, x_i) + \epsilon(x_i),
\]

\[
\mu \sim \text{Tailfree}(\Pi, \mathcal{G}, \Lambda), \quad \epsilon \sim \mathcal{G}\mathcal{P}(0, k_\epsilon).
\]

Here \( \mu : \mathcal{F} \times \mathcal{X} \rightarrow [0, 1] \) is a random measure that controls the contribution of each individual base model \( \hat{f}_k \) to the overall ensemble, depending on the location in the feature space \( x \in \mathcal{X} \). It follows a dependent tail-free process which is constructed from a collection of normalized Gaussian processes \( g \in \mathcal{G} \) and sparse-inducing hyperparameters \( \lambda \in \Lambda \) (See the end of this section for details). The residual process \( \epsilon \) is a flexible Gaussian process that captures the systematic bias shared by the base prediction functions. In terms of uncertainty quantification, \( \mu \) and \( \epsilon \) play distinct roles in the decomposition of the overall uncertainty into that due to model selection and that due to prediction. Specifically, the posterior uncertainty in \( \mu(\hat{f}_k, x) \) reflects uncertainty in model selection, which is expected be high when model predictions disagree and there are few observations to justify confident selection, and is expected to be low otherwise. On the other hand, the posterior uncertainty in \( \epsilon(x) \) reflects uncertainty in prediction, which describes the ensemble’s additional uncertainty in its final prediction, conditional on model selection (See, e.g. Figure 1).

Dependent Tail-free Process

In the simplest scenario, given a set of \( K \) prediction functions \( \mathcal{F} = \{\hat{f}_k\}_{k=1}^K \) with no grouping structure among them, we can consider \( \mathcal{F} \) as a depth 1 tree with \( \mathcal{K} \) leaf nodes (we denote this tree as \( \Pi \)), and model the ensemble weights \( \mu(\hat{f}_k, x) \) as \( P(\hat{f}_k|\text{root}, x) \), i.e. the conditional probability of the \( k \)th leaf node given its parent (i.e. the root node, see Appendix Figure A.2). Specifically, we model \( P(\hat{f}_k|\text{root}, x) \) as the softmax transformation of \( K \) independent Gaussian processes corresponding to each \( \hat{f}_k \):

\[
P(\hat{f}_k|\text{root}, x) = \frac{\exp(g_k(x)/\lambda_{\text{root}})}{\sum_{k'=1}^K \exp(g_{k'}(x)/\lambda_{\text{root}})}, \quad \{g_k\}_{k=1}^K \overset{iid}{\sim} \mathcal{G}\mathcal{P}(0, k_\mu).
\]

Here \( \lambda_{\text{root}} \) is the temperature parameter that controls the sparsity in model selection among the children of \( \text{root} \). Denoting the collection of Gaussian processes \( \{g_k\}_{k=1}^K \) corresponding to each node in the tree \( \Pi \) as \( \mathcal{G} \), and the collection of all temperature parameters in \( \Pi \) as \( \Lambda \), we have specified a dependent tail-free process (DTFP) prior for the ensemble weights \( \mu \sim \text{Tailfree}(\Pi, \mathcal{G}, \Lambda) \)\cite{10}. Figure 1 shows a graphical representation of our model in this simple scenario. In the more common scenario where models exhibit grouping structure, e.g. some \( \hat{f}_k \) in the ensemble
come from the same model family or are trained on the same dataset, we can incorporate this group structure by organizing \( \hat{f}_k \)'s into a corresponding tree structure, and partition the model space \( F \) accordingly (see Appendix Figure A.2). Modelling for \( \mu \) under such recursive partitioning scheme can also be carried out naturally using the tail-free process \([6, 10]\), where the probability of each leaf node is modelled as a sequence of independent conditional probabilities in its ancestry:

\[
\mu(\hat{f}_k, x) = P(\hat{f}_k \mid \text{Parent}(\hat{f}_k), x) \times \left[ \prod_{f \in \text{Anc}(\hat{f}_k)} P(f \mid \text{Parent}(f), x) \right],
\]

where we have denoted \( \text{Anc}(\hat{f}_k) \) as the set of \( \hat{f}_k \)'s ancestors in the tree, and denoted \( \text{Parent}(f) \) the immediate parent of a node \( f \). The conditional probabilities \( P(f \mid \text{Parent}(f), x) \) are modelled similarly as in (1), but with \( g \)'s corresponding to the siblings of \( f \). In this case, the tail-free construction allows differential sparsity in model selection within each model group (by estimating different temperature parameter within each group), and provides information on the importance of the group of models at each level of the hierarchy.

3 Variational Inference

Denoting \( z = \{G, \Lambda, \epsilon, \sigma\} \) as the collection of all model variables, and \( q_\theta(z) \) as the variational family indexed by parameter \( \theta \), our variational objective of interest is the composition of the KL distance between the variational posterior and the model posterior, and the CvM distance between the variational predictive CDF and the empirical CDF of the data, i.e.

\[
L(\theta \mid z, \{x_i, y_i\}) = KL\left[q_\theta(z) \mid\mid p(z \mid x_i, y_i)\right] + \text{CvM}\left[\hat{P}_\theta, y\right],
\]

where \( \hat{P}_\theta(y < t \mid x_i) = \int_{t}^{-\infty} E_{q_\theta(z)}(p(y \mid z, x_i)) dy \) is the variational predictive CDF. To perform variational learning with respect to \( L \), we minimize the KL distance with respect to its negative evidence lower bound (ELBO) following the standard practice \([2]\), and perform Monte Carlo gradient update on the CvM distance. Specifically, we express the CvM distance as the expectation over samples from variational posterior \([8]\):

\[
E(y - y_i) - \frac{1}{2} E((y - y')^2),\]

where \( y, y' \sim E_{q_\theta(z)}(p(y \mid z, x_i)) \), and derive its unbiased gradient estimator using the standard score gradient method \([15, 16]\). Variance reduction techniques such as Rao-Blackwellization \([16]\) can be applied to encourage stable convergence. We use sparse Gaussian process \([19]\) as the variational family for \( G \) and \( \epsilon \), and fully-factored lognormals otherwise. (See Appendix B for detail)

4 Experiment and Application

4.1 Nonlinear function approximation

![Figure 2: Comparison in prediction and uncertainty quantification of different ensemble methods. The grey bands indicating models’ 68%, 95% and 99% predictive interval.](image)

We first investigate the model’s behavior in prediction and uncertainty quantification on a 1-D nonlinear regression task. We compare the performance of our method with five other commonly used ensemble algorithms: (1) \textbf{avg} is the averaging ensemble that simply average over base model predictions, (2) \textbf{cv-stack} is the stacked generalization method \([3]\) that aggregates model using simplex weight by minimizing their combined cross-validation errors. (3) \textbf{lnr-stack} and (4) \textbf{nlr-stack} are the
linear and the nonlinear stacking methods that train a linear regression model or an additive B-spline regression [20] using the base-model predictions as features, and finally (5) gam is the generalized additive ensemble [22] that linearly combines the base-model prediction, and uses an extra smoothing spline term to mitigate any systematic bias. While avg and cv-stack produce only deterministic predictions, lnr-stack, nlr-stack and gam provide predictive distributions for the outcome.

| Model      | Ours     | avg       | cv-stack  |
|------------|----------|-----------|-----------|
| Validation RMSE | 0.1531 ± 0.017 | 0.3723 ± 0.028 | 0.2086 ± 0.014 |

Table 1: Mean and Standard Deviation for validation RMSE in 1-D regression task

Detailed experiment protocol is available in Appendix C.1. Briefly, we generate data from the composition of a global function representing the slow-varying, global trend and a local function representing the fast-varying, local fluctuations. We train four kernel regression models using RBF kernel with length-scale parameters \( \{0.1, 0.2, 0.01, 0.02\} \). As a result, none of the base models can fit the data-generation mechanism universally well across \( x \in (0, 1) \) (see Figure C.4). We train the six ensemble models on seperately generated 20 holdout observations, and evaluate each ensemble model’s RMSE on a validation set of 500 observations. We repeat above training and evaluation procedure 100 times, and report the mean and standard deviation of the validation RMSE in Table 1. We also visualize the models’ behavior in prediction in one such training-evaluation instance in Figure 2. As shown, comparing the mean prediction (blue line), avg, cv-stack and lnr-stack produce either overly smooth or overly complex fits, due to assigning constant weights to the base models. On the other hand, nlr-stack and gam produce closer fit to the data-generation mechanism. But they tend to overfit the observations in the holdout dataset, producing either unnecessary local fluctuations (nlr-stack), or extrapolating improperly in regions outside \( x \in (0, 1) \) (gam), resulting in higher validation RMSE compared to that of our model. In comparison, our model produces smooth fit that closely matches the data-generation mechanism in regions where holdout observation is available or the model agreement is high, and produces smooth interpolation with high uncertainty in regions with few holdout observations and model agreement is low (e.g. \( x \in (0.5, 0.75) \)), indicating proper quantification of uncertainty. Examining the predictive intervals of other ensemble methods, we find that these intervals tend to vary less flexibly within the range of \( x \in (0, 1) \), sometimes failing to reflect the increased uncertainty in regions where the data is sparse and the model agreement is low, and resulting in overly narrow confidence intervals (e.g. gam in \( x \in (0.5, 0.75) \)). We also quantitatively assess these models’ quality in uncertainty quantification (in terms of the true coverage probability of the \( p\% \) predictive intervals for \( p \in (0, 1) \)) in Appendix C.2.

4.2 Spatial integration of air pollution predictions in New England region

In air pollution assessment, many research groups are developing distinct spatio-temporal models (exposure models) to predict ambient air pollution exposures of study participants even in areas where air pollution monitors are sparse. Depending on the prediction model and the inputs, disagreement among model predictions are consistently observed across space and time (Figure D.6), and information on prediction uncertainty are generally unavailable, leading to difficulties in exposure assessment for downstream health effect investigations. Here we use our ensemble method to aggregate the spatiotemporal predictions of three state-of-the-art PM\(_{2.5}\) exposure models ([11, 4, 5] in Figure D.7) to produce a coherent set of spatiotemporal exposure estimate, along with information on predictive uncertainty. We perform our ensemble framework on the base models’ out-of-sample prediction for 43 monitors across greater Boston area during year 2011. We report the ensemble methods leave-one-out RMSE in Table 2 and visualize our model’s posterior prediction and uncertainty in Figure D.8, and the estimated ensemble weights in Appendix Figure D.7. As shown, we observed elevated uncertainty close to Brockton region (where the base models disagrees) and regions further away from metro area (where the monitors are sparse), reflecting uncertainty in model selection and prediction that is consistent with empirical evidence.

| Model      | Ours     | avg       | cv-stack  |
|------------|----------|-----------|-----------|
| loo RMSE   | 0.7580 ± 0.0883 | 1.6768 ± 0.124 | 1.5437 ± 0.1275 |

Table 2: Mean and Standard Deviation for leave-one-out RMSE in annual PM\(_{2.5}\) ensemble prediction
References

[1] T. W. Anderson. On the Distribution of the Two-Sample Cramer-von Mises Criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159, Sept. 1962.

[2] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *arXiv:1601.00670 [cs, stat]*, Jan. 2016. arXiv: 1601.00670.

[3] L. Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, July 1996.

[4] 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.

[5] A. v. Donkelaar, R. V. Martin, M. Brauer, N. C. Hsu, R. A. Kahn, R. C. Levy, A. Lyapustin, A. M. Sayer, and D. M. Winker. Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. Mar. 2016.

[6] T. S. Ferguson. Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics*, 2(4):615–629, July 1974.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] O. Okun, G. Valentini, and M. Re. *Ensembles in Machine Learning Applications*. Springer, Sept. 2011. Google-Books-ID: ZqcMBwAAQBAJ.

[15] J. Paisley, D. Blei, and M. Jordan. Variational Bayesian Inference with Stochastic Search. *arXiv:1206.6430 [cs, stat]*, June 2012. arXiv: 1206.6430.

[16] R. Ranganath, S. Gerrish, and D. M. Blei. Black Box Variational Inference. *arXiv:1401.0118 [cs, stat]*, Dec. 2013. arXiv: 1401.0118.

[17] R. Selten. Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics*, 1(1):43–61, June 1998.
[18] G. Shaddick, M. L. Thomas, A. Green, M. Brauer, A. Donkelaar, R. Burnett, H. H. Chang, A. Cohen, R. V. Dingenen, C. Dora, S. Gumy, Y. Liu, R. Martin, L. A. Waller, J. West, J. V. Zidek, and A. Prüss-Ustün. Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(1):231–253, June 2017.

[19] M. Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Artificial Intelligence and Statistics*, pages 567–574, Apr. 2009.

[20] G. Wahba. *Spline Models for Observational Data*. SIAM, Sept. 1990. Google-Books-ID: ScRQJEETs0EC.

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

[22] 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.
A Additional Figures for Model Description

Figure A.1: Different components of predictive uncertainty decomposed by our model. (left) Ensemble prediction and uncertainty estimate without residual process; (right) Decomposition of uncertainty in model selection (red) and in prediction (grey).

![Ensemble Prediction Mean, Sparse Gaussian Processes](image1)

![Ensemble Prediction Uncertainty Decomposition, Sparse Gaussian Processes](image2)

Figure A.2: An example partition of the model space $F$. In the naive partition on the left, $\mu(f_{model0i}, x) = P(f_{model0i} | root, x)$, and in the right, $\mu(f_{model0i}, x) = P(f_{model0i} | f_{family0}, x)P(f_{family0} | root, x)$

In the right figure, the two conditional probabilities are modelled as:

$$P(f_{model0i} | f_{family0}, x) = \frac{\exp(g_{model0i} / \lambda_{family0})}{\sum_{i=0}^{1} \exp(g_{model0i} / \lambda_{family0})}$$

$$P(f_{family0} | root, x) = \frac{\exp(g_{family0} / \lambda_{root})}{\sum_{i=0}^{1} \exp(g_{family} / \lambda_{root})}$$

(a) A naive partition  
(b) A partition utilizing model family information.
B  Expression for Gradient of the Variational Inference Objective

Variational family Due to our focus on reliable uncertainty quantification, we find the naive mean-field approximation with fully-factored Gaussians tend to under-estimate predictive uncertainty, and produces non-smooth predictions that overfits the observation (see Appendix Figure B.3). Consequently, we adopt a structured approximation based on sparse Gaussian process [19]. Specifically, we factor the variational family into independent groups of Gaussian processes \( \{\epsilon, G\} \) and variance/temperature parameters \( \{\sigma, \Lambda\} \) as

\[
q(z) = q(\epsilon) \ast (\prod_{g \in G} q(g)) \ast \left[ q(\sigma) \ast (\prod_{\lambda \in \Lambda} q(\lambda)) \right],
\]

where we model \( q(\epsilon) \) and \( q(g) \)'s using sparse Gaussian processes, and model \( q(\sigma) \) and \( q(\lambda) \)'s using fully factored log normal distributions.

Figure B.3: Comparison of posterior predictive mean and uncertainty (95% credible interval) between fully-factored mean-field VI (left) and structured VI based on sparse Gaussian process (right).

Gradient of the Variational Inference Objective Denote \( q_\Theta(z) \) the variational distribution and \( p_\Theta(y|x) \) as the variational posterior predictive distribution (i.e. \( p_\Theta(y|x) = \int p(y|z,x) q_\Theta(z) dz \)), then:

\[
\nabla_\Theta KL(q_\Theta \| p) = E_{q_\Theta} \left( (\log p(z,x) - \log q_\Theta(z)) \ast \nabla_\Theta \log q_\Theta(z) \right)
\]

\[
\nabla_\Theta CRPS(q_\Theta, p|x) = E_{p_\Theta} \left( g(y, y_{obs}) \ast \nabla_\Theta \log p_\Theta(y|x) \right) - \frac{1}{2} \left( \nabla_\Theta \log p_\Theta(y|x) + \nabla_\Theta \log p_\Theta(y'|x) \right)
\]

where the gradient for predictive log likelihood \( \log p_\Theta(y|x) \) can be written as:

\[
\nabla_\Theta \log p_\Theta(y) = \frac{E_{q_\Theta}(p(x|z) \ast \nabla_\Theta \log q_\Theta(z))}{E_{q_\Theta}(p(x|z))}
\]
C  Experiment Protocol and Additional Result for the 1-D Regression

C.1 Experiment Protocol

To prepare base models in the ensemble, we randomly generate 20 data points \( x_i \sim \text{Uniform}(0, 1) \), and generate \( y_i = f(x_i + \epsilon_i) \), where \( \epsilon_i \sim \text{i.i.d. N}(0, 0.01) \) and the data-generating function \( f(x) = f_{\text{slow}}(x) + f_{\text{fast}}(x) \) is the composition of a smooth, slow-varying global function \( f_{\text{slow}}(x) = x + \sin(4x) + \sin(13x) \) on \( x \in (-0.5, 1.5) \), and a fast-varying, local function \( f_{\text{fast}}(x) = 0.5 \times \sin(40x) \) on \( x \in (0.1, 0.6) \) (black lines in Figure 2). We train four different kernel regression models on separately generated datasets, using Radial Basis Function (RBF) kernel with two groups of length-scale parameters \( \text{smooth} = \{0.2, 0.1\} \) and \( \text{complex} = \{0.02, 0.01\} \) to represent two groups of models with different smoothness assumptions, resulting in \( \mathcal{F} = \{f_k\}_{k=1} \). As shown in Figure C.4, no base model can predict the ground truth universally well across \( x \in (0, 1) \). We then train all the ensemble methods on a holdout dataset of 20 data points generated using the same mechanism. For our model, we use RBF kernel for both \( \mu \) and \( \epsilon \), where we put prior \( \text{LogNormal}(-1., 1.) \) on the RBF’s length-scale parameters so they are estimated automatically through the inference procedure. The spline hyperparameters for \text{nlr-stack} and \text{gam} are selected using random grid search over \( 10^3 \) candidates based on model’s cross-validation error.

After training, we evaluate each model’s RMSE on a validation dataset of 500 data points spaced evenly between \( x \in (0, 1) \). We repeat above training and evaluation procedure 100 times on randomly generated holdout datasets, and report the mean and standard deviation of the validation RMSE in Table 1. We also visualize the models’ behavior in prediction of one such training-evaluation instance in Figure 2.

Figure C.4: Data generation function and the deterministic predictions from base models in the 1D experiment. Black Line: data-generation function. Colored Line: Base model predictions.

C.2 Quantitative Assessment of Uncertainty Quantification

We also quantitatively assessed \text{gam}, \text{lnr-stack}, \text{nlr-stack} and our model’s quality in uncertainty quantification (in terms of the true coverage probability of the \( p\% \) predictive intervals for \( p \in (0, 1) \)) in Figure C.5. In order to assess the effect of the calibrated VI objective, we compare two versions of our model estimated with and without the CvM distance in VI objective. Compared to other ensemble methods, the nominal coverage of our model’s predictive intervals are shown to be closer to their true coverage, and visible improvement can be observed for the model estimated with the CvM distance included in VI objective.
Figure C.5: Comparison in coverage probability of model’s predictive interval of different ensemble methods. x-axis is the nominal coverage probability of model’s predictive interval (in percentage), and y-axis is the actual coverage probability of model’s predictive interval. Ideally, the coverage curve should align with the black line. For a given coverage percentage, curve below black line indicates underestimated uncertainty (i.e. overly narrow predictive interval), and curve above the black line indicates over-estimated uncertainty (i.e. unnecessarily wide predictive interval).

Top Left: Our model, with KL-only VI objective, Top Right Our model with KL+CvM VI objective. Bottom: (from left to right) gam, lnr-stacking and nlr-stacking.
D  Additional Figures for Real Data Application

Figure D.6: Visualization of annual average $PM_{2.5}$ predictions from different base models in greater Boston region during Year 2011.

Figure D.7: (Top) Base model predictions and (Bottom) the posterior mean of the corresponding ensemble weight Gaussian Process.

Figure D.8: Posterior predictive mean (left) and uncertainty (standard deviation) (right) for annual $PM_{2.5}$ in greater Boston area during year 2011. All unit in $\mu g/m^3$. 

11