Multi-model Mixing via Bayesian Trees

Matthew T. Pratola

Digital Twins for Nuclear & Particle Physics, Trento Italy

web: www.matthewpratola.com
email: mpratola@iu.edu

Tuesday May 12th, 2026

BAND: Bayesian Analysis of Nuclear Dynamics



http://www.bandframework.github.io

- Integrate diverse software tools into a cohesive framework (Taweret).
  - Physics simulators, Emulation, Calibration, Bayesian Model Mixing (BMM)
- BMM Frameworks
  - Moment (mean) mixing
    - Gaussian Process based mixing (Semposki et al.)
    - Tree-based mixing (Yannotty et al.)
  - Density mixing
    - Linear density mixing (Liyanage et al.)
    - Approximate Likelihood mixing (Ingles et al.)

BMM: Bayesian Model Mixing

- How to combine multiple (simulator) models?
- When best model may depend on inputs?
- Or maybe no best model at some or all inputs?
- And quantify uncertainties?

General BMM Framework

- A process, Y, is observed, which depends on some p-dimensional inputs $\mathbf{x}$.
- This observation is of some underlying phenomena of interest, $f_\dagger(\mathbf{x})$.
- A simple model that we often adopt linking the observations to the underlying phenomena is

$$Y|\mathbf{x} \sim N\left(f_\dagger(\mathbf{x}), \mathbf{c}(\cdot)\right)$$

- Problem: we don't know $f_\dagger$.

BMM (Mean Mixing)

- Assume we have K models $\mathcal{M}_1, \ldots, \mathcal{M}_K$. One approach is to average the mean predictions of these models,

$$\hat{f}_l = \hat{E}[Y|\mathcal{M}_l],$$

in our model for Y,

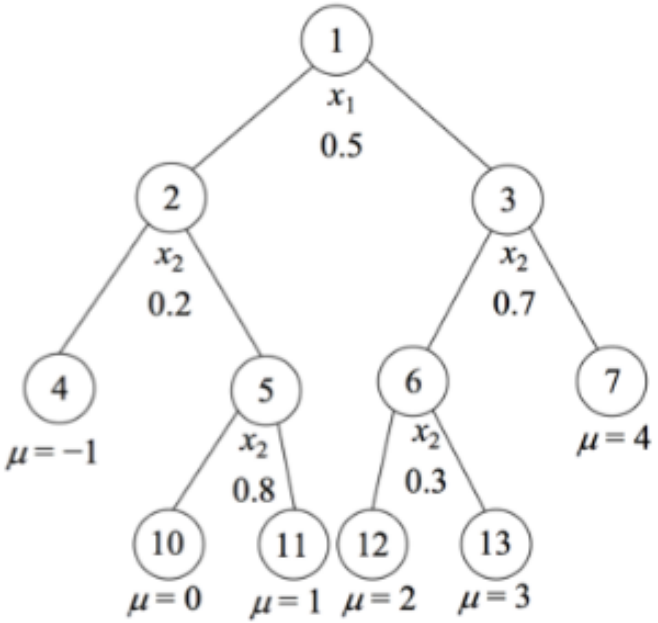
$$Y \sim N\left(\sum_{k=1}^{K} w_k(\mathbf{x})\hat{f}_k(\mathbf{x}), c(\cdot)\right).$$

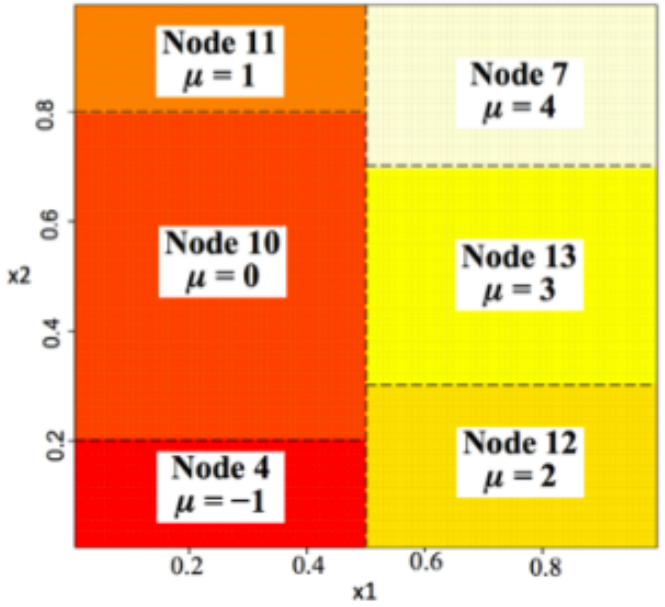
Bayesian Regression Tree Models

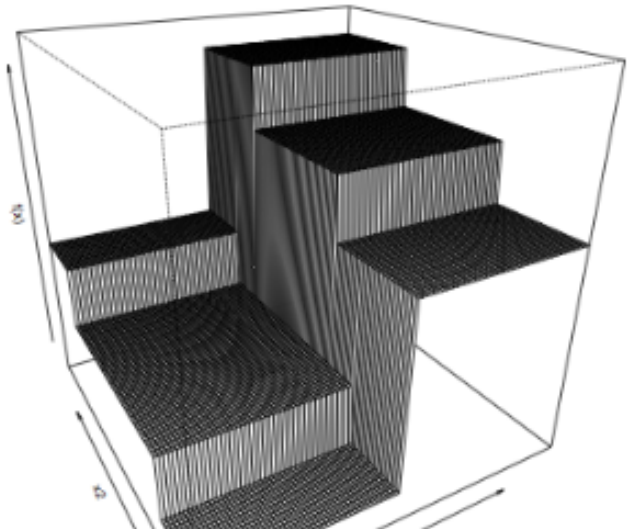
Regression Trees

- Popular tool in Machine/Statistical Learning - divide and conquer!
- Binary trees define "multivariate step function" bases.
- Internal nodes (η's) have `variable<cutpoint` rules ($x_v < c$).
- Coefficients are the terminal node parameters (μ's).
- Each path from terminal node to root defines a basis function.
- $\mathcal{T} = (\{\eta_j\}, \{v_j\}, \{c_j\})$, $\mathcal{M} = (\mu_1, \ldots, \mu_B)$

Regression Trees







Three different views of a bivariate single tree.

Bayesian Additive Regression Trees (BART)

- BART is an additive model:

$$Z(\mathbf{x}) = \sum_{j=1}^{m} g(\mathbf{x}; T_j, M_j) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- Given observations $\mathbf{z} = (z_1, \ldots, z_n)$, we are interested in sampling the posterior distribution

$$\pi(\sigma^2, \{T_j, M_j\}_{j=1}^m | \mathbf{z}) \propto$$

$$L(\sigma^2, \{T_j, M_j\}_{j=1}^m | \mathbf{z}) \pi(\sigma^2) \prod_{j=1}^{m} \pi(M_j | T_j) \pi(T_j)$$

BART-BMM Model

- Conditional on the values of the theoretical predictions at a given point, $f_1(\mathbf{x}_i), \ldots, f_K(\mathbf{x}_i)$, the model can be defined by

$$Y_i \mid \mathbf{f}(\mathbf{x}_i), \mathbf{w}(\mathbf{x}_i), \sigma^2 \sim N(\mathbf{f}^\top(\mathbf{x}_i)\mathbf{w}(\mathbf{x}_i), \sigma^2)$$

where $\mathbf{f}(\cdot) = (f_1(\cdot), ..., f_K(\cdot))^\top$ and $\mathbf{w}(\cdot) = (w_1(\cdot), ..., w_K(\cdot))^\top$.

- The weight vectors are modeled as a sum-of-trees,

$$\mathbf{w}(\mathbf{x}_i) = \sum_{j=1}^{m} \mathbf{g}(\mathbf{x}_i, T_j, M_j, Z_j)$$

where $\mathbf{g}(\mathbf{x}_i, T_j, M_j, Z_j)$ is the K-dimensional output of the jth tree using the set of terminal node parameters, M_j, at the input, $\mathbf{x}_i$

Yannotty et al., *Model Mixing using Bayesian Additive Regression Trees*, Technometrics, vol. 66 (2024)

Terminal Node Prior

- For the pth terminal in tree j,
  $\mu_{pj} \mid T_j \sim_{i.i.d.} N_K(\frac{1}{mK}\mathbf{1}_K, \tau^2\mathbf{I}_K)$.

- The induced prior on the lth model weight is
  $w_l(\mathbf{x}_i) \sim N(\frac{1}{K}, m\tau^2)$.

- Then τ selected so that $w_l(\mathbf{x}_i) \in [0, 1]$ with high probability. To do so, a CI for $w_l(\mathbf{x}_i)$ is constructed s.t.

$$0 = 0.5 - k\tau\sqrt{m} \quad \text{and} \quad 1 = 0.5 + k\tau\sqrt{m}.$$

  Subtracting the first equation from the second and solving yields

$$\tau = \frac{1}{2k\sqrt{m}}.$$

- Default choice is $k = 1$. Larger k concentrates on $1/K$, smaller k allows more flexibility.

Tree-depth Prior

- Exponentially decreasing prior probability that the ith internal node η_{ji} of tree $\mathcal{T}_j$ will split:

$$\pi(\eta_{ji} \text{ splits}) = \alpha(1 + d(\eta_{ji}, \eta_{j1}))^{-\beta}$$

  where $d(\eta_{ji}, \eta_{j1})$ is the depth from node η_{ji} to the root node η_{j1} in tree $\mathcal{T}_j$.

- Defaults are $\alpha = 0.95, \beta = 2$ implies trees of depth 2-3.

EFT Expansions and why not just use Bayesian Model Averaging (BMA)?

Effective Field Theories

- Effective Field Theories (EFT's) are a modeling framework popular in Nuclear Physics.
- Allow approximating behavior of particles using an expansion that captures increasingly fine structed details of the physics.
- "Think" T-S expansion, except expansion is by order of physical effects.
- Typically approximate real-world behavior by truncating the expansion to a computationally amenable order. In previous work, under a GP prior we formulate the distribution of both the retained expansion and the residual[†].

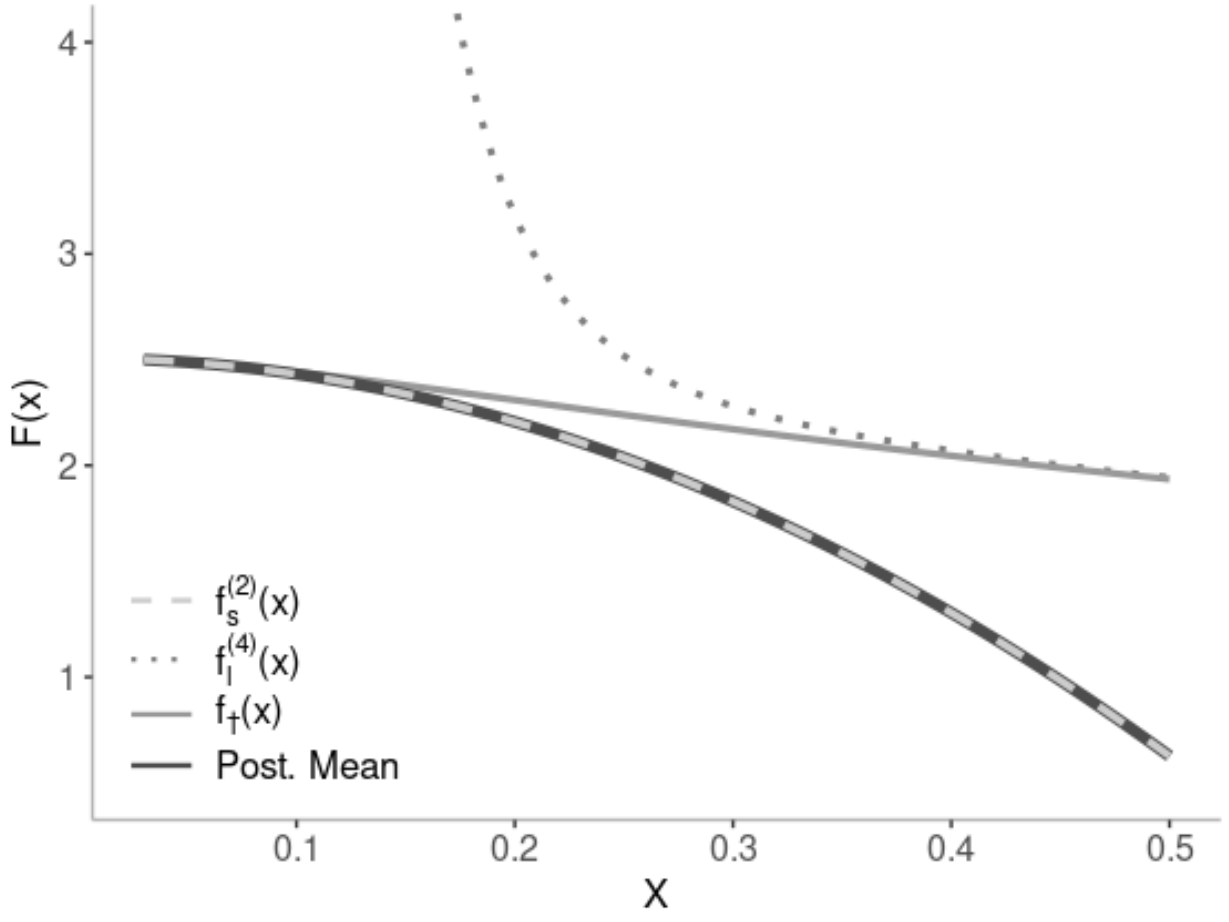
[†] Melendez et al. (2019)

BMA

- Weighting of models has a long history, e.g. Bayesian Model Averaging (BMA):

$$p(f_\dagger) = \sum_{k=1}^{K} w_k p(f_\dagger | \mathcal{M}_k)$$

BMA



Motivating EFT from Honda (2014). The posterior mean prediction of $f_\dagger(x)$ when applying BMA to the 2nd order weak and 4th order strong coupling expansions.

BMA

- Weighting of models has a long history, e.g. Bayesian Model Averaging (BMA):

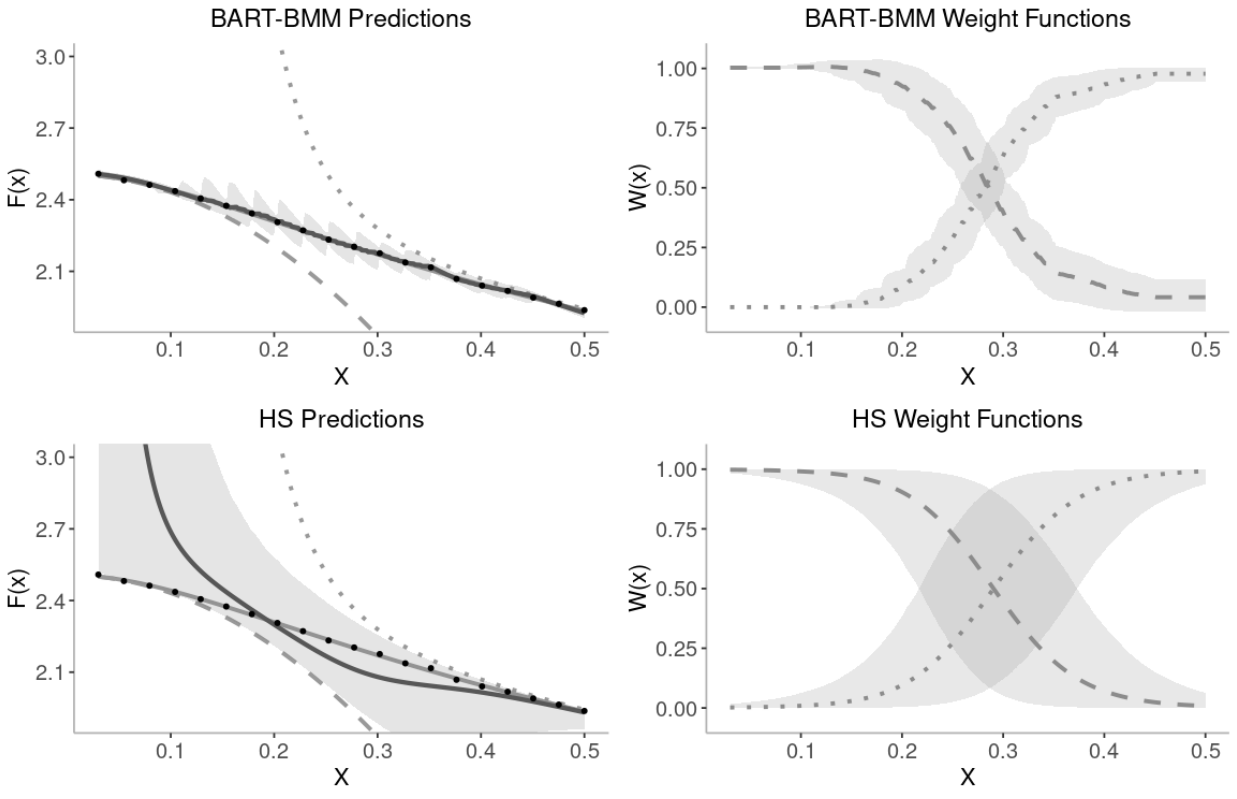
$$p(f_\dagger) = \sum_{k=1}^{K} w_k \, p(f_\dagger | \mathcal{M}_k)$$

- Strong assumptions of global weights and M-closed setting.

Effective Field Theories

- As it turns out, there are usually many EFT expansions that are indexed by some variable (energy is one possibility).
- Multiple EFTs models available in different regimes of the input domain – great locally, weak globally.
- Given a collection of K such models, would like to combine them in some effective way.
- Existing tools not appropriate: BMA (global); Stacking (global); Hierarchical Stacking (user specifies mean)

EFT Expansion Example



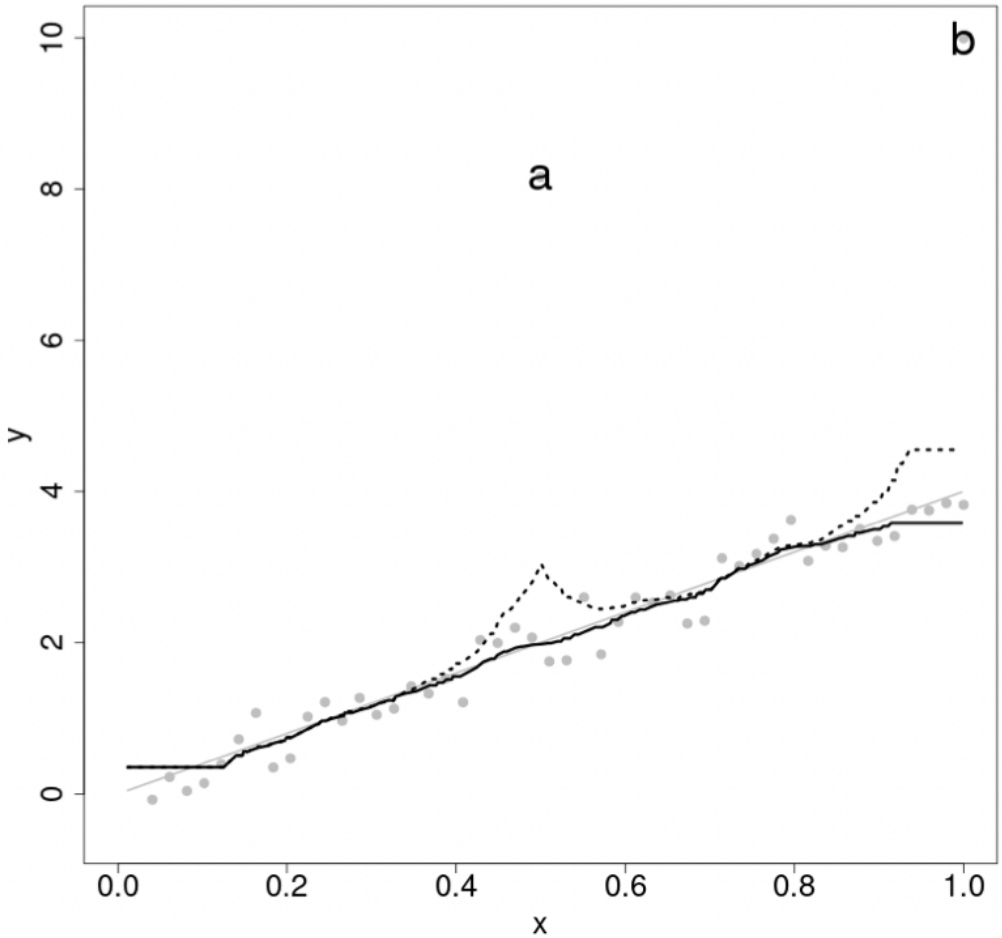
The predicted mean (dark gray) and 95% credible intervals (shaded) when mixing $f_s^{(2)}(x)$ (dashed) and $f_l^{(4)}(x)$ (dotted). Results are obtained from a BART-BMM model with 10 trees and a Hierarchical Stacking model with a linear unconstrained weight function (bottom).

Smooth BART-based Model Mixing

Higher-Dimensional BMM

- No obvious reason why this can't extend well into higher-dimensional problems.
- But in practice we have observed overfitting issues.
- This problem arises in regular BART regressions as well.

E.g. BART Overfitting



The dotted line is the BART prediction with outliers A and B included in the data, the black line is the BART prediction with the outliers removed.

Pratola et al., *Influential Observations in Bayesian Regression Tree Models*, Journal of Computational and Graphical Statistics, vol. 33 (2024)

Random Path BART (RPBART)

- Follow left/right splits at internal nodes with some probability.
- Latent indicator variable $z_{bj}(\mathbf{x}_i)$ denotes if observation i maps to terminal node b in tree j.
- Definining $Z_j = \{\mathbf{z}_j(\mathbf{x}_i)\}_{i=1}^n$ where $\sum_{b=1}^{B_j} z_{bj}(\mathbf{x}_i) = 1$ we have

$$Y(\mathbf{x}_i)|\{T_j, M_j, Z_j\}_{j=1}^m, \sigma^2 \sim N\left(\sum_{j=1}^m g(\mathbf{x}_i; T_j, M_j, Z_j), \sigma^2\right)$$

where

$$g(\mathbf{x}_i; T_j, M_j, Z_j) = \sum_{b=1}^{B_j} \mu_{bj} z_{bj}(\mathbf{x}_i).$$

Yannotty et al., *Bayesian Model Mixing of Computer Simulators with Applications to Climate*, Journal of Machine Learning Research (2026)

Random Path BART (RPBART)

- Idea: conditional on Z_j's we still have usual BART step functions. But marginally we have a smooth, continuous function.
- Prior:

$$\mathbf{z}_j(\mathbf{x}_i)|\, T_j, \gamma_j \sim \text{Multinomial}(1, \phi_{1j}(\mathbf{x}_i; T_j, \gamma_j), \ldots, \phi_{B_j j}(\mathbf{x}_i); T_j, \gamma_j)$$

- $\phi_{bj}()$ is the probability an observation maps to terminal b in tree j.
- $\gamma_j \in (0, 1)$ is a bandwidth parameter, with prior

$$\gamma_j \sim \text{Beta}(\alpha_1, \alpha_2), j = 1, \ldots, m.$$

Path Probabilities

- The path probabilities are determined by the probability of branching left/right at each internal node along paths from root node to termninal nodes:

$$\phi_{bj}(\mathbf{x};\, T_j, \gamma_j) = \prod_{d=1}^{D} \psi(\mathbf{x};\, v_{(dj)}, c_{(dj)}, \gamma_j)^{R_{(dj)}} \times \left(1 - \psi(\mathbf{x};\, v_{(dj)}, c_{(dj)}, \gamma_j)\right)^{1-R_{(dj)}}$$

where $R_{(dj)} = 1$ ($R_{(dj)} = 0$) denotes if a right (left) move defines the path and

$$\psi(\mathbf{x};\, v_{(dj)}, c_{(dj)}, \gamma_j) = \begin{cases} 1 - \frac{1}{2}\left(1 - \frac{x_{v_{(dj)}} - c_{(dj)}}{\gamma_j(U_v^d - c_{(dj)})}\right)_+^q & x_{v_{(dj)}} \geq c_{(dj)}, \\ \frac{1}{2}\left(1 - \frac{c_{(dj)} - x_{v_{(dj)}}}{\gamma_j(c_{(dj)} - L_v^d)}\right)_+^q & x_{v_{(dj)}} < c_{(dj)}, \end{cases}$$

Smooth Predictions

- Marginalizing over the random path assignments gives us the desired smooth fitted response,

$$E[Y(\mathbf{x}) \mid \{T_j, M_j, \gamma_j\}_{j=1}^{m}] = \sum_{j=1}^{m} \sum_{b=1}^{B_j} \boldsymbol{\mu}_{bj}\, \phi_{bj}(\mathbf{x}; T_j, \gamma_j).$$

- How do we calibrate the priors? Not enough time in this talk...

RPBART-BMM Model

- Conditional on the values of the theoretical predictions at a given point, $f_1(\mathbf{x}_i), \ldots, f_K(\mathbf{x}_i)$, the model can be defined by

$$Y_i \mid \mathbf{f}(\mathbf{x}_i), \mathbf{w}(\mathbf{x}_i), \sigma^2 \sim N\big(\mathbf{f}^\top(\mathbf{x}_i)\mathbf{w}(\mathbf{x}_i), \sigma^2\big)$$

  where $\mathbf{f}(\cdot) = \big(f_1(\cdot), ..., f_K(\cdot)\big)^\top$ and $\mathbf{w}(\cdot) = \big(w_1(\cdot), ..., w_K(\cdot)\big)^\top$.

- The weight vectors are modeled as a vectorized RPBART,

$$\mathbf{w}(\mathbf{x}_i) = \sum_{j=1}^{m} \mathbf{g}(\mathbf{x}_i, T_j, M_j, Z_j)$$

  where $\mathbf{g}(\mathbf{x}_i, T_j, M_j, Z_j)$ is the K-dimensional output of the jth tree using the set of terminal node parameters, M_j, at the input, $\mathbf{x}_i$

Yannotty et al., *Bayesian Model Mixing of Computer Simulators with Applications to Climate*, Journal of Machine Learning Research (2026)

Post-Processing

- We post-process the fitted model weights to decompose into a simplex weight and a discrepancy term:

$$E[Y(\mathbf{x})] = \sum_k v_k(\mathbf{x}) f_k(\mathbf{x}) + \delta(\mathbf{x})$$

where $v_k(\mathbf{x}) \in [0,1]^d$ and $\sum_k v_k(\mathbf{x}) = 1, \forall \mathbf{x}$.

- Provides interpretation of which simulator works well where, and what patterns the simulators miss (discrepancy).

Global Climate Model (GCM) Example

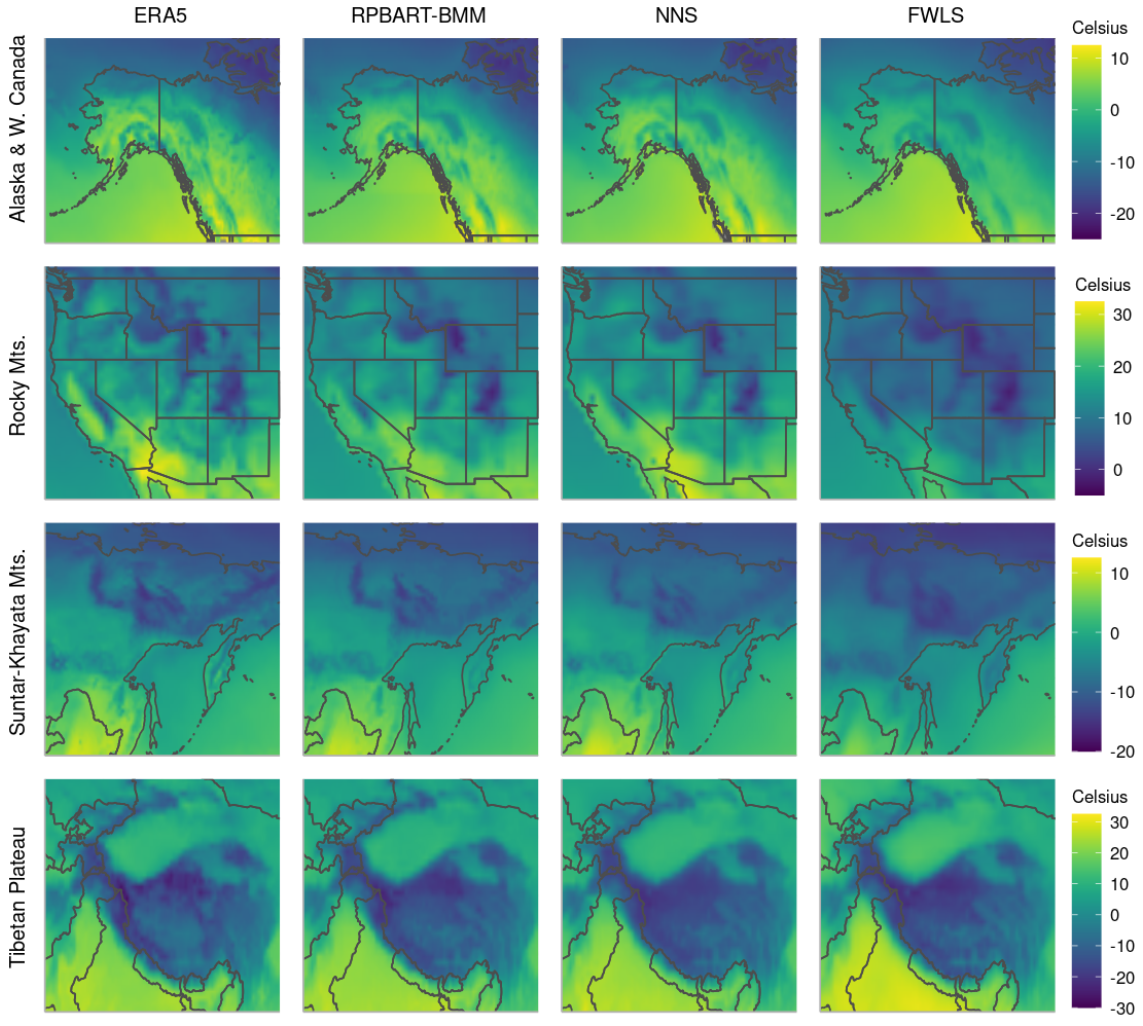
Climate Models

- Motivated by work of Harris et al.[†] – empirically observed some models seemed better over land, others over water.
- Combining multiple GCM's has been a topic of interest for some time (e.g Knutti et al.[⋆]).
- Apply BMM to perform local mixing of a set of 8 climate models.

[†] Harris et al., *Multimodel ensemble analysis with neural network Gaussian processes*, The Annals of Applied Statistics, vol.17 (2023)

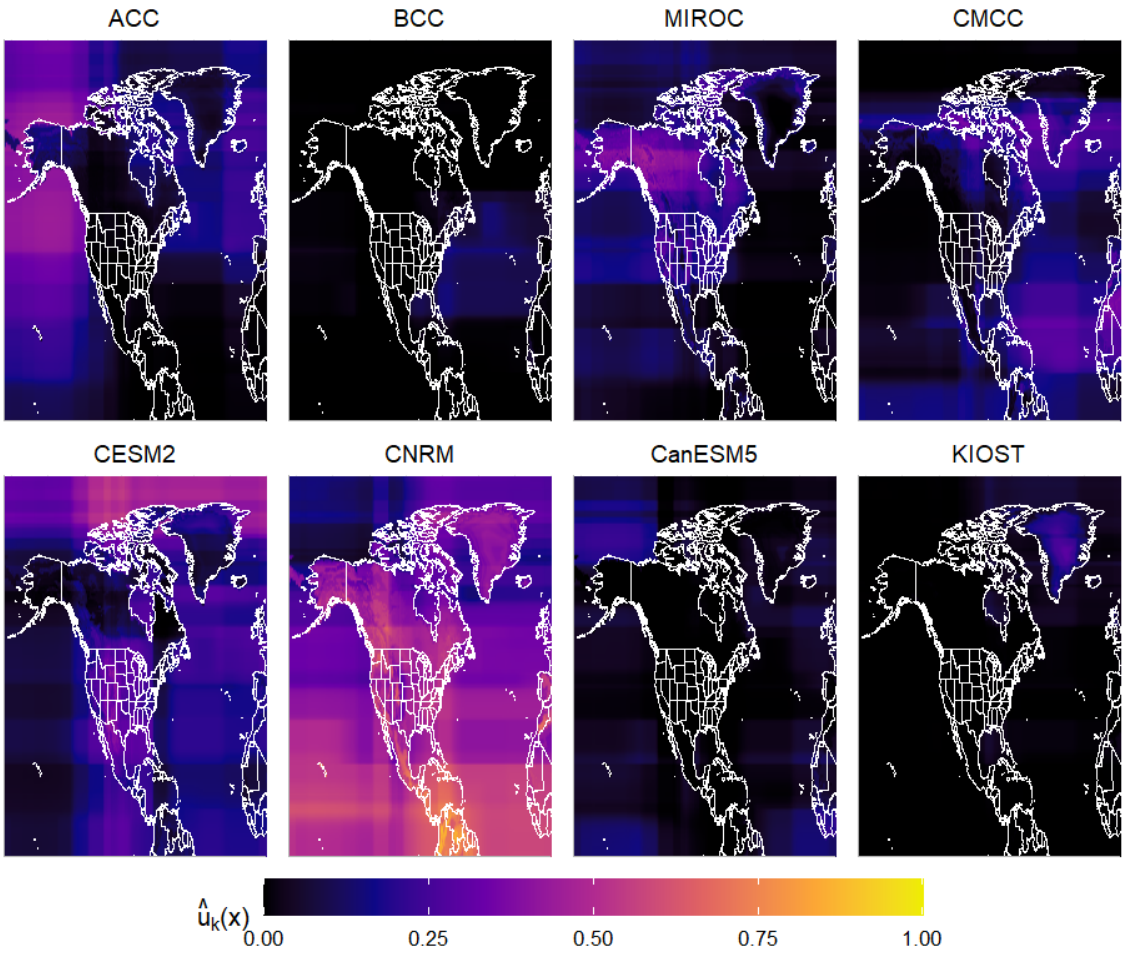
[⋆] Knutti et al., *Challenges in combining projections from multiple climate models*, Journal of Climate, vol.23 (2010).

Climate Models Example



RPBART-BMM fit to ERA5 data using 8 climate simulators, compared to three stacking alternatives.

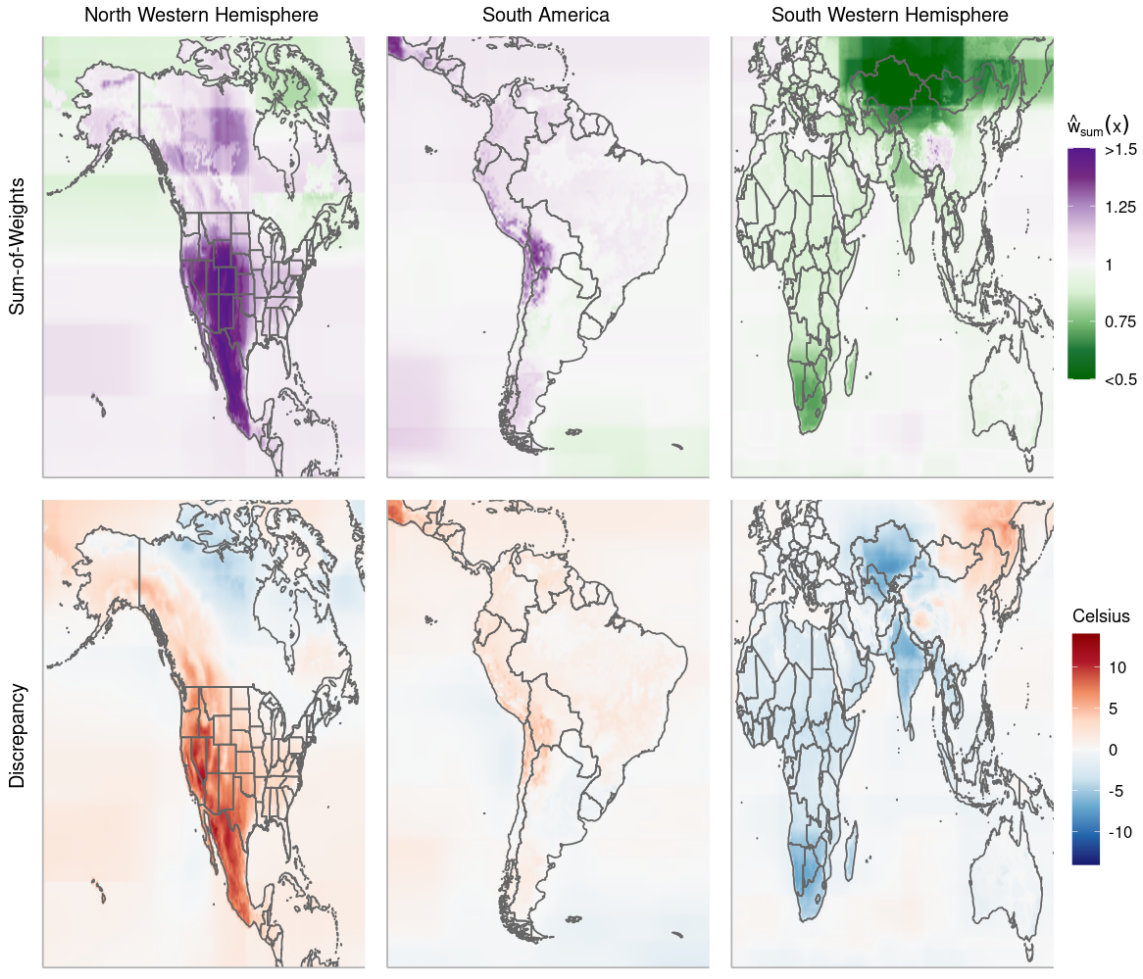
Climate Models Example



Projected constrained weights of Laha et al.[†] minimizing the discrepancy.

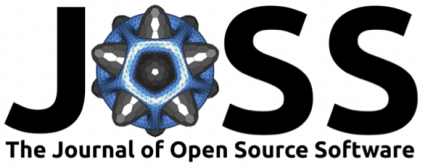
[†] Laha et al., *On controllable sparse alternatives to softmax*, Advances in Neural Information Processing Systems, vol.31 (2018)

Climate Models Example



Sum of unconstrained weights (top row) and discrepancy remaining after constrained prediction of ERA5 field using the constrained weights (bottom row).

Taweret



Taweret: a Python package for Bayesian model mixing

K. Ingles [1*¶], **D. Liyanage** [2*], **A. C. Semposki** [3*], **and J. C. Yannotty** [4*]

1 Illinois Center for Advanced Study of the Universe & Department of Physics, University of Illinois Urbana-Champaign, USA **2** Department of Physics, The Ohio State University, USA **3** Department of Physics and Astronomy & Institute of Nuclear and Particle Physics, Ohio University, USA **4** Department of Statistics, The Ohio State Univeristy, USA ¶ Corresponding author * These authors contributed equally.

Conclusion

- BART is a highly successful divide and conquer strategy.
- BART-BMM for (mean) mixing multiple simulation models of a phenomena of interest.
- RPBART-BMM adds smoothness with a principled approach to prior calibration.
- Posterior projections for interpretation.
- General procedure for combining multiple predictive models where one expects the local best model to vary. Digital twins?

Conclusion

- Want to learn Bayesian tree models?

  http://www.matthewpratola.com/teaching/stat8810-fall-2017/

  (slides 11-14)