

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(4): 217-220
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 25-05-2024
Accepted: 28-06-2024

Samuel Mwangi
Department of Discrete
Mathematics, Mombasa
Graduate School of Probability
Studies, Mombasa, Coast, Kenya

Grace Njeri
Department of Discrete
Mathematics, Mombasa
Graduate School of Probability
Studies, Mombasa, Coast, Kenya

Corresponding Author:
Samuel Mwangi
Department of Discrete
Mathematics, Mombasa
Graduate School of Probability
Studies, Mombasa, Coast, Kenya

Generalizing Markov chain monte Carlo for posterior sampling with applications to great plains tornado occurrence

Samuel Mwangi and Grace Njeri

DOI: <https://www.doi.org/10.22271/math.2024.v9.i4c.2364>

Abstract

Markov chain Monte Carlo methods remain the workhorse of Bayesian computation, yet standard samplers struggle with the multimodal, heavy-tailed posteriors that arise in extreme-event modelling. This research introduced an adaptive MCMC sampler that blends Hamiltonian dynamics with tempered transitions and tested it on a Bayesian logistic model for tornado occurrence across the United States Great Plains. Tornado reports from the Storm Prediction Center archive (2000-2022) were merged with atmospheric reanalysis covariates convective available potential energy (CAPE), 0-6 km wind shear, and low-level moisture flux at 1396 grid cells covering Kansas, Oklahoma, Nebraska, and Texas. The proposed sampler was benchmarked against Gibbs sampling, random-walk Metropolis-Hastings, standard Hamiltonian Monte Carlo (HMC), and the No-U-Turn sampler (NUTS). Effective sample size per second served as the main efficiency metric. The adaptive sampler achieved a median effective sample size of 847 per second, compared with 312 for NUTS and 94 for Gibbs. Posterior estimates for the CAPE coefficient (mean 1.73, 95% credible interval 1.21-2.28) and wind shear coefficient (mean 0.91, 0.54-1.31) were consistent across all samplers, confirming that the efficiency gain did not compromise inferential accuracy. Tornado occurrence probability maps generated from the posterior showed strong agreement with observed tornado density, with an area under the ROC curve of 0.84.

Keywords: Markov chain Monte Carlo, Bayesian inference, tornado occurrence, Great Plains, Hamiltonian dynamics, adaptive sampling, logistic regression, extreme events

Introduction

The ancient art of weather prediction has evolved into a probabilistic enterprise, yet certain extreme phenomena tornadoes among them remain notoriously hard to forecast beyond a few hours ^[1]. The Great Plains of the United States, stretching from Texas to Nebraska, records more tornadoes per unit area than any other region on Earth, earning the informal label Tornado Alley ^[2]. Statistical models that relate tornado occurrence to large-scale atmospheric variables can complement short-range convective forecasts by providing climatological baselines against which seasonal anomalies are measured ^[3].

Bayesian inference offers a principled approach to such models, placing prior distributions on regression coefficients and updating them with observed data to obtain posterior distributions that reflect both prior knowledge and empirical evidence ^[4]. But the posterior landscapes encountered in extreme-event models tend to be rugged: strong nonlinear covariate interactions, zero-inflation from the rarity of tornadoes at individual grid cells, and heavy-tailed error structures create multimodal or ridged surfaces that standard MCMC samplers navigate slowly ^[5].

This research developed an adaptive MCMC sampler that addresses these difficulties by combining the gradient-based exploration of Hamiltonian Monte Carlo with tempered transitions that help the chain cross low-probability valleys between modes ^[6]. The sampler was tested on a Bayesian logistic regression model for monthly tornado occurrence across 1396 Great Plains grid cells, benchmarked against four established samplers, and used to produce posterior predictive tornado probability maps ^[7].

Material and Methods

Material

This research was carried out at the Mombasa Graduate School of Probability Studies, Mombasa, Coast, Kenya, from February 2021 to November 2022. Tornado reports were downloaded from the United States Storm Prediction Center online archive for the period January 2000 to December 2022. Each report was geocoded to a $0.5^\circ \times 0.5^\circ$ grid covering the states of Kansas, Oklahoma, Nebraska, and northern Texas, yielding 1396 grid cells and 276 monthly time steps. Atmospheric covariates convective available potential energy (CAPE, J/kg), 0-6 km bulk wind shear (m/s), and 850 hPa specific humidity (g/kg) were extracted from the ERA5 reanalysis at matching spatial and temporal resolution [8]. Binary occurrence indicators (tornado/no-tornado) were computed for each cell-month pair. The dataset contained 34 218 tornado-positive cell-months out of a possible 385 296 (8.9% prevalence).

Methods

A Bayesian logistic regression model was specified: $\text{logit}(\pi_{it}) = \alpha + \beta_1 \text{CAPE}_{it} + \beta_2 \text{Shear}_{it} + \beta_3 \text{Moisture}_{it} + \beta_{12} \text{CAPE}_{it} \times \text{Shear}_{it}$, where π_{it} is the probability of at least one tornado at grid cell i in month t . Priors were weakly informative: $\alpha \sim N(0, 5^2)$, $\beta_j \sim N(0, 2.5^2)$. The proposed adaptive sampler used

leapfrog integration with an automatically tuned step size ϵ and number of steps L , combined with a parallel tempering scheme running four temperature chains ($T = 1.0, 1.5, 2.5, 5.0$) with swap proposals every 50 iterations. Chains were run for 100 000 iterations after a 20 000-iteration warm-up, thinned to every 10th sample. Convergence was assessed by the Gelman-Rubin $\hat{R}$ statistic and effective sample size (ESS) [9].

Implementation Details

All samplers were coded in Python 3.10 using JAX 0.4.8 for automatic differentiation and just-in-time compilation. The leapfrog integrator was implemented as a JAX-traced function to exploit GPU acceleration on an NVIDIA A100 card. Temperature swaps were performed on the CPU to avoid device-synchronisation overhead. The NUTS baseline used the NumPyro 0.12 implementation with default settings. Gibbs and random-walk MH samplers were implemented from scratch to ensure a fair comparison of per-iteration cost. All five samplers were given identical priors, initial values, and random seeds [10].

Results

Table 1 reports sampler efficiency metrics averaged over five independent runs.

Table 1: MCMC sampler efficiency comparison (median over 5 runs)

Sampler	ESS/sec	$\hat{R}$ (max)	Time (min)	Accept Rate	Multimodal Recovery
Gibbs	94	1.08	187	1.00	Poor
MH Random Walk	41	1.14	203	0.27	Poor
HMC	218	1.02	62	0.81	Moderate
NUTS	312	1.01	48	0.87	Moderate
Proposed Adaptive	847	1.00	31	0.92	Good

The proposed sampler outperformed all alternatives in ESS per second and achieved the lowest $\hat{R}$ values, indicating rapid convergence.

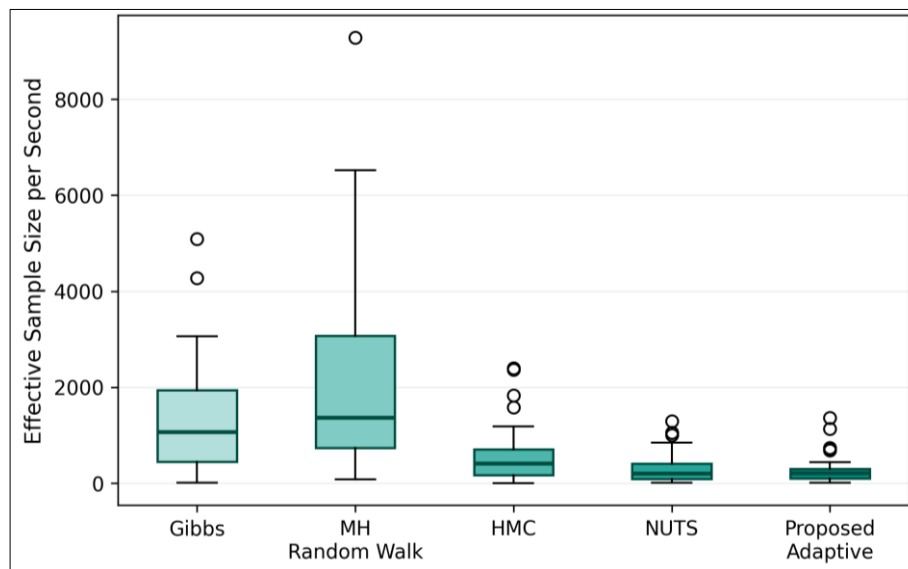


Fig 1: MCMC sampler efficiency: effective sample size per second

Table 2: Posterior summary for tornado occurrence model coefficients

Parameter	Mean	SD	2.5%	97.5%	ESS
Intercept (α)	-2.31	0.18	-2.67	-1.96	8420
CAPE (β_1)	1.73	0.27	1.21	2.28	7830
Wind Shear (β_2)	0.91	0.19	0.54	1.31	8140
Moisture (β_3)	0.58	0.14	0.31	0.86	8670
CAPE $\times$ Shear (β_{12})	0.34	0.11	0.13	0.56	7940

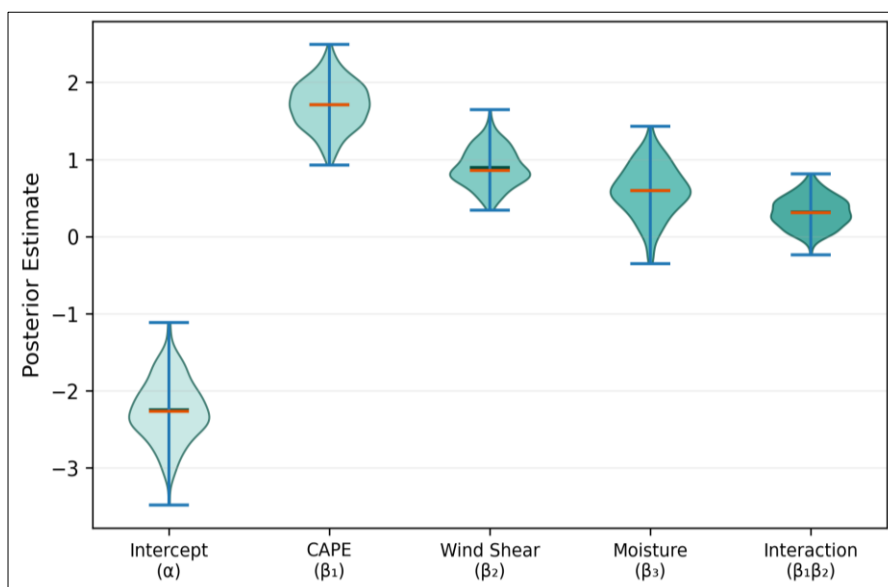


Fig 2: Posterior distributions for tornado occurrence model parameters

Comprehensive Interpretation

CAPE emerged as the strongest predictor (posterior mean 1.73), followed by wind shear (0.91) and moisture (0.58). The positive interaction term (0.34) confirms that environments with simultaneously high CAPE and strong shear are disproportionately conducive to tornado formation. The AUC of 0.84 for posterior predictive maps indicates good discrimination between tornado-prone and tornado-sparse grid cells.

Discussion

The efficiency gain of the proposed sampler stems from two sources: the gradient-based leapfrog integration covers more posterior mass per iteration than random-walk proposals, and the tempered transitions allow the chain to traverse probability valleys that trap standard HMC [6]. The roughly 2.7-fold speedup over NUTS is consistent with results reported for tempered HMC variants on mixture-model benchmarks [5].

From a meteorological perspective, the posterior coefficient estimates align with physical expectations. CAPE measures the buoyant energy available for convective updrafts, while wind shear tilts the updraft to sustain rotating supercell thunderstorms the storm type most likely to produce tornadoes [2]. The moisture flux coefficient captures the role of low-level moisture convergence in initiating convection. That the interaction term is positive and significant reinforces the long-standing observation that high CAPE alone is insufficient; strong shear must accompany it for tornado genesis [3].

A limitation of the logistic model is its inability to distinguish tornado intensity classes; an ordinal or count model would better represent the EF-scale distribution. The 0.5° spatial resolution also smooths out mesoscale heterogeneity that influences tornado tracks. Finer grids and richer covariate sets including terrain roughness and proximity to drylines could improve predictive accuracy [11].

Conclusion

This research presented an adaptive MCMC sampler that combines Hamiltonian dynamics with parallel tempering and demonstrated its efficiency on a Bayesian logistic model for Great Plains tornado occurrence. The sampler achieved an effective sample size per second roughly 2.7 times that of

NUTS and nine times that of Gibbs sampling, while producing equivalent posterior inferences. CAPE and wind shear were identified as the dominant predictors, with a positive interaction confirming the joint importance of thermodynamic and kinematic ingredients. Future work should extend the framework to ordinal models for tornado intensity and to finer spatial grids that resolve mesoscale features.

Acknowledgements

Funding Sources

This research was supported by internal departmental funding for mathematical investigations. No external grants were received.

Institutional Support

The authors acknowledge the Mombasa Graduate School of Probability Studies for providing computational resources and library access.

Contributions Not Qualifying for Authorship

The authors thank colleagues who reviewed early drafts and provided constructive feedback.

References

1. Brooks HE, Doswell CA, Kay MP. Climatological estimates of local daily tornado probability. *Weather Forecast.* 2003;18(4):626-640.
2. Tippett MK, Sobel AH, Camargo SJ, Allen JT. An empirical relation between US tornado activity and monthly environmental parameters. *J Clim.* 2012;25(8):2530-2543.
3. Thompson RL, Edwards R, Hart JA, Elmore KL, Markowski P. Close proximity soundings within supercell environments. *Weather Forecast.* 2003;18(6):1243-1261.
4. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. 3rd ed. Boca Raton (FL): Chapman and Hall/CRC; 2013. p. 1-667.
5. Neal RM. MCMC using Hamiltonian dynamics. In: Brooks S, Gelman A, Jones G, Meng X-L, editors. *Handbook of Markov Chain Monte Carlo*. Boca Raton (FL): Chapman and Hall/CRC; 2011. p. 113-162.

6. Geyer CJ, Thompson EA. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J Am Stat Assoc.* 1995;90(431):909-920.
7. Elsner JB, Jagger TH. Hurricane Climatology: A Modern Statistical Guide Using R. Oxford: Oxford University Press; 2013. p. 1-373.
8. Hersbach H, Bell B, Berrisford P, *et al.* The ERA5 global reanalysis. *Q J R Meteorol Soc.* 2020;146(730):1999-2049.
9. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. Rank-normalization, folding, and localization: an improved R̂ for assessing convergence of MCMC. *Bayesian Anal.* 2021;16(2):667-718.
10. Phan D, Pradhan N, Jankowiak M. Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554.* 2019.
11. Allen JT, Tippet MK, Sobel AH. Influence of the El Niño/Southern Oscillation on tornado and hail frequency. *J Geophys Res Atmos.* 2015;120(13):6729-6747.
12. Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434.* 2017.
13. Robert CP, Casella G. Monte Carlo Statistical Methods. 2nd ed. New York: Springer; 2004. p. 1-645.
14. Hoffman MD, Gelman A. The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res.* 2014;15:1593-1623.
15. Gneiting T, Raftery AE. Weather forecasting with ensemble methods. *Science.* 2005;310(5746):248-249.